



Etude d'une famille multigénique chez *Encephalitozoon cuniculi*, une microsporidie pathogène de l'homme

Ndongo Dia

► To cite this version:

Ndongo Dia. Etude d'une famille multigénique chez *Encephalitozoon cuniculi*, une microsporidie pathogène de l'homme. Biologie du développement. Université Blaise Pascal - Clermont-Ferrand II; Université d'Auvergne - Clermont-Ferrand I, 2006. Français. NNT : 2006CLF21724 . tel-00713572

HAL Id: tel-00713572

<https://theses.hal.science/tel-00713572>

Submitted on 2 Jul 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

UNIVERSITE BLAISE PASCAL

UNIVERSITE D'Auvergne

N° D.U : 1724

**Ecole Doctorale
des Sciences de la Vie et de la Santé**
N° Ordre : 450

THESE

pour l'obtention du grade de

DOCTEUR D'UNIVERSITE

spécialité : **Parasitologie Moléculaire**

présentée et soutenue publiquement par

Ndongo DIA

le 18 Décembre 2006

ETUDE D'UNE FAMILLE MULTIGENIQUE CHEZ *ENCEPHALITOZON CUNICULI*, UNE MICROSPORIDIE PATHOGENE DE L'HOMME

JURY

Président : Pr. B. S. TOGUEBAYE, Université Cheikh Anta Diop (UCAD) de DAKAR
Pr. Christian VIVARES, Université Blaise Pascal, Clermont-Ferrand.

Rapporteurs : Frédéric BRINGAUD, Directeur de recherche Laboratoire de Genomique
Fonctionnelle des Trypanosomatides UMR-CNRS 5162
Serge CASAREGOLA, Directeur de recherche INRA CNRS INA-PG
UMR Microbiologie et Génétique moléculaire.

Directeur de thèse : Pr. Emmanuel CORNILLLOT, Université Blaise Pascal, Clermont-Ferrand.

*Laboratoire de Biologie des Protistes UMR CNRS 6023, équipe de Parasitologie moléculaire et cellulaire.
Université Blaise pascal, Clermont-Ferrand*

Tous les remerciements vont vers Toi...

Remerciements

Ce travail a été réalisé au sein du laboratoire de Biologie des Protistes à l'université Blaise Pascal. Mes profonds remerciements à Monsieur **Christian Amblard**, Directeur du laboratoire. A travers vous je remercie tout le personnel du laboratoire.

Mes sincères et profonds remerciements aux Professeurs **Christian Vivares** et **Ben S. Toguebaye** qui sont à l'origine de ce travail. Vous m'avez permis de découvrir la biologie moléculaire. Veuillez trouver ici l'assurance de mon estime et de toute ma reconnaissance.

Je tiens à remercier particulièrement **Emmanuel Cornillot** pour son encadrement, ses conseils et son soutien permanent tout au long de ce travail. Tu m'as initié à la biologie moléculaire et particulièrement à la génomique. Ta rigueur scientifique m'a beaucoup marquée. En tout cas sois assuré de ma sympathie et de mon estime.

Je souhaite exprimer mes profonds remerciements à **Frédéric Bringaud** et à **Serge Casaregola**. Vous m'honorez en acceptant de juger ce travail malgré vos nombreuses responsabilités.

J'exprime ma sincère reconnaissance à Monsieur **Guy Méténier**. Votre gentillesse, votre disponibilité et votre sens élevé des relations humaines me touchent beaucoup. Soyez assuré de mon estime.

Mes sincères remerciements à **Laurence Lavie** et à **Ngor Faye**. A vos côtés, j'ai beaucoup appris de la biologie moléculaire. Trouvez ici l'assurance de ma grande reconnaissance.

Je remercie **Gerrard Prensier, Catherine Texier, Frédéric Delbac, Hicham El-Alaoui**, vous m'avez toujours proposé vos services. Soyez assurés de ma reconnaissance.

Mes remerciements à **Valérie, Ivan, Damien, Muriel, Vanessa, Marie, Samuel**...bref toute l'équipe de parasitologie. Soyez assurés de ma sympathie. Je n'oublie pas mon frère et ami, **Lamjed Mansour**.

A ma mère...

à toute la famille...

... à tous les amis

TABLE DES MATIERES

Remerciements

TABLE DES MATIERES	1
INTRODUCTION	5
ETUDE BIBLIOGRAPHIQUE	8

I- Le génome : définition, structure et complexité	9
II- La chromatine	12
II-1 Le nucléosome, unité fondamentale de la chromatine	12
II-2 Les protéines chromatinienne non-histones	14
II-3 Les états de condensation de la chromatine	15
II-3-1 Le complexe condensine	17
II-3-2 Le complexe cohésine	19
III- Le chromosome : éléments généraux de structure et de fonction	19
III-1 Origines de réplication	19
III-2 Centromères	21
III-2-1 Centromères ponctuels	21
III-2-2 Centromères régionaux	24
III-3 Extrémités chromosomiques	25
III-3-1 Télomères	26
III-3-1-1 Régulation de la longueur des télomères	28
III-3-1-2 Protection contre les recombinaisons	33
III-3-1-3 Regroupement des télomères dans le noyau	
et « effet de position »	35
III-3-2 Subtélomères	37
III-3-2-1 Organisation	37
III-3-2-2 Importance dans la recombinaison homologue	39

IV- Les petits génomes nucléaires chez les Eucaryotes unicellulaires	41
IV-1 Séquençage systématique et nécessité de la cartographie	42
IV-1-1-1 Cartographie par électrophorèse à champs pulsés	43
IV-1-1-2 Fragmentation aléatoire et agencement clonal	46
IV-2 Exemples de petits génomes	47
IV-2-1 Chez les Protistes	47
- <u>Apicomplexes</u>	
- <u>Kinétoplastidés</u>	
- <u>Diplomonadines et Trichomonadines</u>	
- <u>Entamoeba histolytica</u>	
IV-2-2 Chez des Champignons unicellulaires	49
- <u>Hémiascomycètes</u>	
- <u>Archiascomycètes</u>	
IV-2-3. Chez Dictyostelium discoideum (Mycetozoa)	50
V- Familles multigéniques impliquées dans la variation antigénique chez les parasites	51
V-1 Exemples de familles multigéniques	52
V.1.1. Famille <i>var</i> chez <i>Plasmodium falciparum</i>	52
V.1.2. Famille <i>vsg</i> chez <i>Trypanosoma brucei</i>	55
V.1.3. Famille <i>vsp</i> chez <i>Giardia intestinalis (lamblia)</i>	58
V.1.4. Exemples de familles chez des champignons	59
V-2 Régulation de l'expression des familles multigéniques	60
V-2-1 Promoteurs et sites d'expression	60
V-2-1-1 Gènes <i>var</i> : trois types de promoteurs	60
V-2-1-2 Gènes <i>vsg</i> : plusieurs sites d'expression avec un promoteur de type ribosomique	62
V-2-1-3 Gènes <i>msg</i> : un seul site d'expression incluant une région transcrite commune à tous les ARNm	64
V-2-2 L'exclusion allélique	65
V-2-2-1 Contrôle par recombinaison de l'ADN	65
V-2-2-2 Activation <i>in situ</i>	68
V-2-3 Contrôle épigénétique	70

V- Les microsporidies	74
V-1 Généralités	74
V-2 Position phylogénétique : un changement radical	76
V-3 Structure de la spore	77
V-4 Germination et invasion cellulaire	79
V-5 Développement intracellulaire	81
V-6 Diversité des cycles biologiques	84
V-7 Génome nucléaire	86
V-7-1 Caryotype moléculaire	86
V-7-1-1. Variations interspécifiques	86
V-7-1-2 Variations intraspécifiques	88
V-7-1-3 Cartographie en 2D (par la technique KARD-PFGE)	89
V-7-2 Données du séquençage systématique	89
V-7-2-1 Chez <i>Encephalitozoon cuniculi</i>	89
V-7-2-2 Chez <i>Antonospora locustae</i>	91
V-7-2-3 Chez <i>Vittaforma corneae</i>	91

RESULTATS	92
------------------	-----------

CHAPITRE 1 :	93
---------------------	-----------

Régionalisation et organisation mosaïque des subtélomères des chromosomes de la microsporidie *Encephalitozoon cuniculi*. (*Article 1: Subtelomeres in the microsporidian *Encephalitozoon cuniculi*: regionalization and mosaic organization*).

CHAPITRE 2:	151
--------------------	------------

La famille multigénique *interB*: un répertoire de gènes associé aux subtélomères chez le parasite *Encephalitozoon cuniculi*. (*Article 2: InterB multigenic family, a gene repertoire associated with subterminal chromosome regions of *Encephalitozoon cuniculi* and conserved in several human-infecting microsporidian species*).

CHAPITRE 3: Expression des protéines InterB	169
I- Expression de protéines InterB recombinantes	169
I-1 Production de protéines recombinantes	172
I-2 Obtention des anticorps polyclonaux	174
I-3 Extraction et séparation des protéines sporales	174
I-4 Transfert sur membrane et immunomarquage	175
I-5 Immunolocalisation	175
II- Résultats	175
II-1 Réactivité des anticorps anti-Y111_ENCCU	175
II-2 Expression de différentes régions de newECU03 chez <i>E. coli</i>	176
 Discussion et perspectives	 179
<hr/>	
I- Le répertoire <i>interB</i> chez <i>Encephalitozoon cuniculi</i>	180
II- Transcription des gènes <i>interB</i> : quel contrôle?	181
III- Protéines InterB et localisation cellulaire	184
IV- Gènes <i>interB</i> : relation avec la spécificité d'hôte?	185
<hr/>	
 Références bibliographiques	 187

Introduction

Introduction :

Le travail de cartographie physique des 11 chromosomes qui composent le génome de la microsporidie *Encephalitozoon cuniculi* avait révélé la présence d'une unité ADN_r (23S-16S) à chaque extrémité d'un chromosome. Trois unités sont présentes sur les 11 consensus représentatifs des 11 chromosomes, indiquant clairement un assemblage incomplet des régions télomériques et subtélomériques. L'utilisation du logiciel Miropeat pour comparer les séquences chromosomiques entre elles révèle la présence de séquences homologues aux différentes extrémités. Au total 15 répétitions (r01 à r15) ont été mises en évidence. Ces dernières sont organisées en 8 blocs nommés EXT1 à EXT8 composant ainsi les différentes extrémités des chromosomes. Cette structure en mosaïque des extrémités a été confirmée par une approche PCR. Par cette approche toutes les extrémités chromosomiques ont été entièrement reconstruites (avec un ancrage sur l'ADN_r sur toutes les extrémités). Par rapport à la dynamique de ces régions, 19 sites de recombinaisons ont été retrouvés en utilisant l'algorithme Blast. Ces régions montrent également un biais compositionnel en bases GC. En effet l'entrée dans les séquences terminales (blocs EXT et unité ADN_r) est systématiquement marquée par une augmentation brutale du taux de GC. Par contre il n'existe pas de corrélation entre les taux (G-C)/(G+C) ou (A-T)/(A+T) et l'entrée dans les régions EXT. Ces résultats montrent une accumulation d'événements au niveau de ces régions particulières des chromosomes que nous avons appelées « régions subterminales ». Elles sont situées entre l'unité ADN_r et le cœur codant des chromosomes. Nous retrouvons la même régionalisation des extrémités chez le nucléomorphe *Guillardia theta*. Ceci suppose que cette organisation est liée au processus d'évolution régressive du génome. Le plus intéressant est la présence de gènes dans ces régions. En effet l'analyse des séquences montrent la présence de séquences codantes (révélées par l'annotation) au niveau des différents blocs EXT. La comparaison de ces séquences codantes révèle 4 familles de gènes : *interAE*, *interB*, *interC* et *interD*. Ces gènes de par leur contexte, répondent à un type d'organisation classique que nous retrouvons chez un certain nombre de protozoaires parasites (gènes *var* chez *Plasmodium*, *vsg* chez les trypanosomes, *msg* chez *Pneumocystis*...), d'où une potentielle implication dans la pathogénicité de ce parasite. Nous avons choisi d'investir la famille *interB*, une famille homogène, avec très peu de gènes incomplets comparée aux trois autres familles. Après la validation de la présence des gènes *interB* chez les espèces du genre *Encephalitozoon*, ces gènes ont été retrouvés chez deux autres genres de microsporidies pouvant parasiter l'homme

(*Brachiola algerae* et *Vittaforma corneae*). Par contre, ces gènes sont absents chez 5 autres genres de microsporidies, dont 2 parasites stricts d'insectes (*Nosema bombycis* et *Paranosema grylli*) et 3 parasites stricts de poissons (*Glugea atherinae*, *Spraguea lophii* et *Neonosemoides syacii*). L'expression de ces gènes a été abordée chez l'espèce *Encephalitozoon cuniculi*. Les gènes *interB* sont transcrits, et des protéines *InterB* existent également. L'étude de ces gènes reste à compléter par une localisation des protéines correspondantes au niveau du parasite (ou même au niveau de la cellule hôte).

Etude bibliographique

Etude bibliographique

I- Le génome : définition, structure et complexité

Le **génom** représente l'ensemble de l'information héréditaire des organismes vivants. Il contient toutes les instructions nécessaires au développement, au fonctionnement, au maintien de l'intégrité et à la reproduction de toute entité biologique (l'exception à la règle étant les prions). Ces instructions, nommées **gènes**, ne constituent qu'une partie du génome (pour exemple, le génome humain coderait pour 25 000 gènes seulement, ce qui ne représente que 5% du génome ; Stein, 2004).

Chez les virus, le génome est constitué soit d'une molécule d'ADN simple ou double brin, soit d'une molécule d'ARN à polarité positive (dans le même sens qu'un ARN messager) ou négative (complémentaire d'un ARN messager).

Par contre, chez les organismes cellulaires, le matériel génétique est exclusivement représenté par l'ADN double brin et différents génomes peuvent coexister.

Chez les **Procaryotes**, on distingue habituellement :

- le génom chromosomique, génome principal se présentant classiquement sous forme d'un seul chromosome circulaire.
- le génom extrachromosomique, constitué par les plasmides, molécules plus petites et à répllication autonome; certains plasmides sont capables d'intégration dans le chromosome (épisomes).

En fait, deux chromosomes différents ou plus ont été mis en évidence chez des bactéries de divers genres (*Vibrio*, *Brucella*, *Deinococcus*, *Rhizobium*...). En outre, il existe des plasmides linéaires et l'ADN chromosomique est parfois aussi sous forme linéaire (*Borrelia*, *Streptomyces*, *Rhodococcus*...) (revue Bendich et Drlica, 2000)

Chez les **Eucaryotes**, trois types de génomes sont distingués en fonction de leur localisation intracellulaire :

- le génom nucléaire, génome majoritaire composé de plusieurs chromosomes linéaires et localisé dans le noyau, structure ayant primaiement servi à définir les cellules eucaryotes (*eu* : vrai, *caryos* : noyau). Les cycles biologiques avec reproduction sexuée impliquent l'alternance haploïde-diploïde. Une fraction d'ADN extrachromosomique (épisomes et éléments transposables) est également retrouvée dans le noyau.
- le génom mitochondrial (dans les mitochondries), d'origine endosymbiotique et souvent sous forme circulaire. Sa taille est particulièrement réduite chez les Vertébrés (16,5

kb chez l'Homme ; Anderson et al, 1981). Cet ADN est linéaire chez certains Protistes (*Tetrahymena*, *Chlamydomonas*, *Pythium*) et certains Champignons (*Candida*, *Pichia*, *Williopsis*).

- le génom chloroplastique (dans les chloroplastes), autre génome extra-nucléaire d'origine endosymbiotique, caractéristique des cellules photosynthétiques. Sa taille se situe typiquement entre 120 et 200 kb chez les Végétaux supérieurs. Des formes circulaires et linéaires co-existent fréquemment.

Bien qu'il y ait chevauchement partiel des tailles génomiques entre Virus, Procaryotes et Eucaryotes, il est évident que ce sont les Eucaryotes qui présentent les génomes les plus complexes, d'où l'utilisation d'unités Mb ou « millions de paires de bases » (**Tableau 1**). Constante pour une espèce déterminée, la valeur C correspond à la taille du génome nucléaire haploïde. Parmi les valeurs C relativement faibles, on peut noter celles de la levure *Saccharomyces cerevisiae* (12,5 Mb), du nématode *Caenorhabditis elegans* (100 Mb) et de la mouche *Drosophila melanogaster* (160-180 Mb). Chez les mammifères, la taille génomique peut être de l'ordre de 3 000 Mb. Un extrêmement grand génome est celui du protozoaire *Amoeba dubia* : 670 000 Mb ! (Friz, 1968). La valeur C varie parfois très fortement parmi les espèces d'un même taxon. Au sein des métazoaires, les Amphibiens représentent la classe la plus variable (700-90 000 Mb). Chez les plantes, les Angiospermes ont des génomes allant de 50 Mb à 125 000 Mb et les Ptéridophytes de 98 Mb à 307 000 Mb. Il n'y a donc pas de stricte corrélation avec le degré de complexité de l'organisme, en considérant tant la diversité de ses types cellulaires que le nombre total de ses gènes. C'est ce qu'on appelle le « paradoxe de la valeur C » (revue : Novacek et Norell, 1989). Le nombre de gènes par génome varie effectivement dans une moindre mesure : ~6000 chez *S. cerevisiae*, ~13 000 chez la drosophile, ~25 000 chez l'Homme...On sait désormais que la disproportion entre taille génomique et nombre de gènes résulte d'une accumulation plus ou moins importante de séquences d'ADN non codant, de type répétitif, dont les fonctions sont plutôt obscures. L'énigme persistante est donc le pourquoi de la grande variabilité de l'ADN non codant chez des espèces aux complexités similaires.

Agée seulement d'une dizaine d'années, la **génomique** est l'étude exhaustive des génomes. Elle implique notamment la détermination des séquences complètes, la détection de tous les gènes potentiels et la recherche des fonctions géniques. Son essor a été intimement lié au développement du séquençage automatisé et d'outils informatiques pour l'assemblage de multiples blocs de séquence (contigs), la délimitation des gènes et la recherche d'homologies.

Espèce	Groupe taxonomique	Taille du génome (Mb)	Références
<u>VIRUS</u>			
<i>Virus de l'hépatite B</i>	Hepadnavirus	0,0032*	Okamoto et al. 1986
<i>Bactériophage λ</i>	Siphovirus	0,0485*	Sanger et al. 1982
<i>Mimivirus APM</i>	Mimivirus	1,18*	Raoult et al. 2004
<u>PROCARYOTES</u>			
<i>Nanoarchaeum equitans</i>	Archaea	0,49*	Waters et al. 2003
<i>Sulfolobus solfataricus</i>	Archaea	2,99*	She et al. 2001
<i>Methanosarcina acetivorans</i>	Archaea	5,75*	Galagan et al. 2002
<i>Mycoplasma genitalium</i>	Bacteria	0,58*	Fraser et al. 1995
<i>Escherichia coli</i> K12	Bacteria	4,64*	Blattner et al. 1997
<i>Pseudomonas fluorescens</i>	Bacteria	7,07*	Paulsen et al. 2005
<u>EUCARYOTES</u>			
<i>Giardia intestinalis</i>	Diplomonadines	12	Adam 2000
<i>Trichomonas vaginalis</i>	Trichomonadines	180	Upcroft et al. 2006
<i>Bigelowiella natans</i>	Chlorarachniophytes	0,38* (nucléomorphe)	Gilson et al. 2006
<i>Guillardia theta</i>	Cryptophytes	0,55* (nucléomorphe)	Douglas et al. 2001
<i>Ostreococcus tauri</i>	Chlorophytes	12,5*	Derelle et al. 2006
<i>Leishmania major</i>	Euglenozoaires	33	Myler et al. 2000
<i>Trypanosoma brucei</i>	Euglenozoaires	39	El-Sayed et al. 2000
<i>Thalassiosira pseudonana</i>	Diatomées	34*	Armbrust et al. 2004
<i>Coscinodiscus asteromphalus</i>	Diatomées	25 000	Shuter et al. 1983
<i>Cryptosporidium hominis</i>	Apicomplexes	9,2*	Xu et al. 2004
<i>Plasmodium falciparum</i>	Apicomplexes	23*	Gardner et al. 2002
<i>Toxoplasma gondii</i>	Apicomplexes	87	Blaxter and Ivens 1999
<i>Tetrahymena thermophila</i>	Ciliés	104* (macronoyau) 200 (micronoyau)	Eisen et al. 2006
<i>Heterocapsa pygmaea</i>	Dinoflagellés	4 450	Triplett et al. 1993
<i>Gonyaulax polyedra</i>	Dinoflagellés	98 000	Shuter et al. 1983
<i>Entamoeba histolytica</i>	Amoebozoaires	24*	Loftus et al. 2005
<i>Dictyostelium discoideum</i>	Amoebozoaires	34*	Glockner et al. 2002
<i>Amoeba dubia</i>	Amoebozoaires	670 000	Friz 1968
<i>Encephalitozoon cuniculi</i>	Microsporidies	2,9*	Katinka et al. 2001
<i>Glugea atherinae</i>	Microsporidies	19,5	Bidre et al. 1994
<i>Saccharomyces cerevisiae</i>	Ascomycètes	12,5*	Goffeau et al. 1996
<i>Neurospora crassa</i>	Ascomycètes	43*	Schulte et al. 2002
<i>Arabidopsis thaliana</i>	Streptophytes	125*	Arabidopsis Genome Initiative 2000
<i>Oryza sativa</i>	Streptophytes	466*	Yu et al. 2002
<i>Zea mays</i>	Streptophytes	3 000	Arumuganathan and Earle 1991
<i>Triticum aestivum</i>	Streptophytes	16 000	Consortium wheat genome 2005
<i>Drosophila melanogaster</i>	Insectes	137*	Adams et al. 2000
<i>Mus musculus</i>	Mammifères	2 500*	Waterston et al. 2002
<i>Homo sapiens</i>	Mammifères	2 900*	Waterston et al. 2002

Tableau 1 : Taille des génomes dans différents groupes d'organismes (les astérisques marquent les génomes entièrement séquencés).

L'assemblage des contigs est d'autant plus difficile que le génome étudié est plus grand et riche en séquences répétées. Ainsi, les premiers génomes séquencés furent de petits génomes viraux, mitochondriaux et chloroplastiques. L'année 1995 a fait date car marquée

par l'obtention des premières séquences complètes de chromosomes procaryotes, à savoir le chromosome d'*Haemophilus influenzae* (1830 kb; Fleischmann et al, 1995) puis celui de *Mycoplasma genitalium* (580 kb ; Fraser et al, 1995). L'année suivante a vu l'achèvement du séquençage des 16 chromosomes constituant le génome nucléaire de la levure *Saccharomyces cerevisiae* (12,5 Mb ; Goffeau et al, 1996). Le projet « Génome Humain » est aussi arrivé à son terme (2850 Mb, soit 99% du génome; Lander et al, 2001 ; International Human Genome Sequencing Consortium, 2004). A ce jour (13-10-2006), 383 génomes procaryotes et 21 génomes eucaryotes ont été séquencés (Cf. statistiques au site <http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>). De nombreux projets sont en cours de réalisation.

II- La chromatine

De la comparaison de la longueur totale de l'ADN génomique avec la taille d'un nucléoïde ou d'un noyau, il ressort la nécessité d'une forte condensation de l'ADN. Par exemple, chez l'espèce humaine, près de deux mètres d'ADN doivent être concentrés dans un noyau diploïde dont le diamètre est seulement de quelques μm . La compaction du génome implique différents niveaux de surenroulement de l'ADN et, dans le cas du génome nucléaire des Eucaryotes, elle dépend d'étroites interactions entre l'ADN et certaines protéines. Ce complexe nucléoprotéique dans les trois dimensions de l'espace correspond à la **chromatine**, terme cytologique primairement proposé pour désigner un réseau basophile dans le noyau interphasique.

II-1 Le nucléosome, unité fondamentale de la chromatine

A son plus faible degré de condensation, la chromatine est sous forme d'une fibre de 10 nm ou nucléofilament, qui apparaît comme une succession d'unités similaires régulièrement espacées, les **nucléosomes**. Chaque nucléosome comprend un court segment d'ADN régulièrement associé à huit molécules de protéines histones. Deux régions sont distinguées : la particule cœur et la région de liaison (ou région internucléosomale) reliant les particules cœur adjacentes. La particule cœur, dont la structure est très conservée parmi les espèces, contient un segment d'ADN de longueur constante (146 pb) qui s'enroule (~ 1,7 tour) autour d'un octamère protéique comprenant deux exemplaires de chacune des histones nucléosomales H2A, H2B, H3 et H4 (Kornberg RD, 1974) (**Figure 1**). La longueur de l'ADN

de liaison varie entre 10 et 60 pb selon l'espèce et le type cellulaire. C'est au niveau de cette région qu'une histone internucléosomale (famille H1/H5) est incorporée.

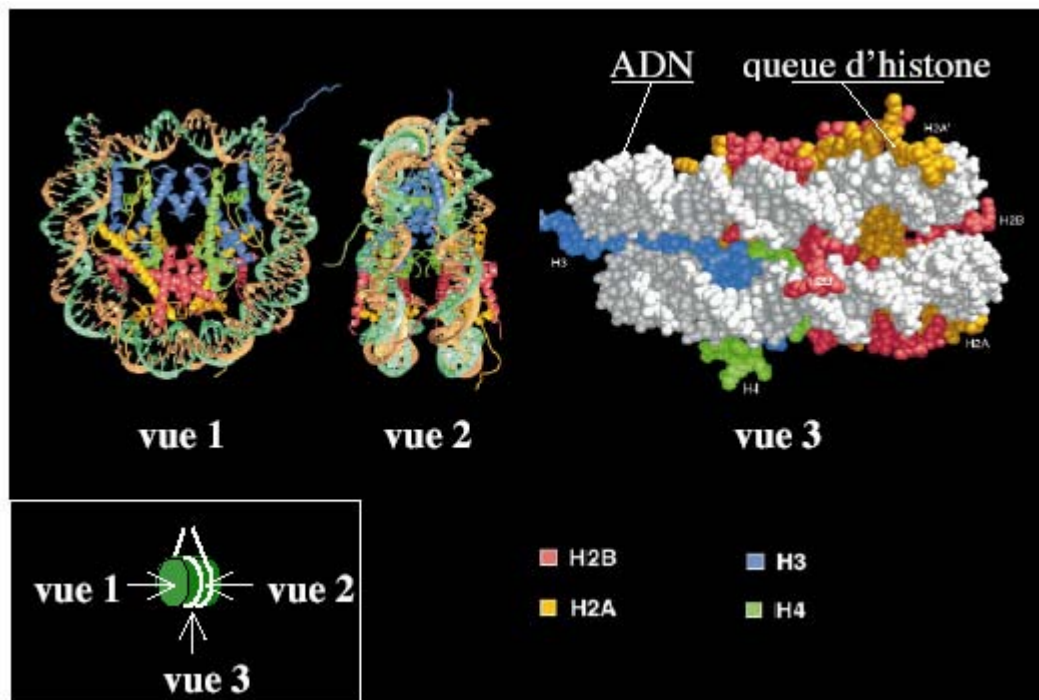


Figure 1: Structure du nucléosome. (D'après *Luger et al, 1997*).

La double hélice d'ADN s'enroule autour de l'octamère d'histones comprenant deux exemplaires de chacune des histones H2A, H2B, H3 et H4.

Les quatre histones de la particule cœur sont de petites protéines basiques dont les séquences ont été très conservées au cours de l'évolution. Elles présentent un même domaine structural (« histone fold ») constitué d'une grande hélice alpha centrale ($\alpha 2$) entourée de deux petites hélices alpha ($\alpha 1$ et $\alpha 3$) ainsi que de boucles (L1 et L2) réalisant des contacts avec l'ADN. Plus sujettes à variation, les extrémités N- et C-terminales n'ont pas de structure secondaire bien définie. Ces extrémités sont particulièrement riches en résidus lysine et arginine et sont donc très basiques. Elles sont la cible de nombreuses modifications post-traductionnelles pouvant affecter leurs charges mais aussi l'accessibilité à l'ADN et les interactions protéine-protéine.

Contrairement aux précédentes, les histones de la famille H1/H5 sont peu conservées parmi les espèces (Baxeavanis et Landsman, 1996). Chez les Eucaryotes supérieurs, elles comprennent un domaine globulaire central non polaire, essentiel pour les interactions avec

l'ADN, et deux extrémités non structurées, hautement basiques, et soumises à des modifications post-traductionnelles. L'association d'un nucléosome avec une histone H1/H5 est appelée **chromatosome** (Simpson R., 1978). Les histones internucléosomales participent à la compaction de l'ADN en créant une région d'interaction entre les nucléosomes adjacents (Ridgway et Almouzni, 2001).

Des liaisons hydrogènes et ioniques permettent au complexe histones-ADN d'être extrêmement stable. Il est important de mentionner aussi que les liaisons entre histones et ADN n'impliquent aucune séquence particulière de l'ADN, ceci étant en accord avec le fait que l'octamère d'histones est virtuellement capable d'empaqueter n'importe quelle séquence (Luger et al, 1997).

II-2 Les protéines chromatinienne non-histones

Outre les histones internucléosomales, des protéines chromatinienne non-histones appelées **HMG** (**H**igh **M**obility **G**roup) et des variants d'histones affectent l'espacement et le repliement du nucléofilament.

Les HMG sont des protéines de faible poids moléculaire, qui correspondent à des facteurs architecturaux de la chromatine bien distincts des histones (Bustin, 1999). Elles sont capables de modifier la conformation de la double hélice pour faciliter les différentes activités cellulaires liées à l'ADN (transcription, réplication, recombinaison, réparation...).

Les variants d'histones semblent jouer des rôles spécifiques dans la structure ou la fonction de la chromatine. C'est le cas notamment d'un variant de H3, nommé CENP-A (CENtromérique Protein A) chez les Eucaryotes supérieurs et Cse4 chez la levure *S. cerevisiae*. Ce variant est associé aux nucléosomes exclusivement au niveau des centromères. Il possède une extrémité N-terminale très différente de H3 et joue un rôle essentiel pour la structure et la fonction des nucléosomes centromériques (Henikoff et al, 2000). Un autre variant de H3, nommé H3.3, a été caractérisé chez la drosophile. Associé à la chromatine active, H3.3 serait essentiel à la régulation de certains gènes, notamment le rDNA (Ahmad et Henikoff, 2002). H2A possède au moins trois variants: H2A-Z, H2A-X et MacroH2A. La protéine H2A-Z est requise pour la formation de l'euchromatine en empêchant, lorsqu'elle est incorporée, la propagation de l'hétérochromatine (Meneghini et al, 2003). Il est à noter que la structure cristalline du nucléosome contenant ce variant est très peu modifiée par rapport à un octamère classique (Suto et al, 2000). H2A-X possède une extrémité C-terminale très différente de celle de H2A chez les métazoaires. Celle-ci peut être spécifiquement phosphorylée en réponse à certains types de cassures dans l'ADN (Rogakou et al, 1998) et

elle est donc requise pour la réparation de l'ADN. Caractérisé par une extrémité C-terminale globulaire de 25 kDa, le variant Macro-H2A est spécifiquement incorporé dans le chromosome X inactif chez les Mammifères (revue : Ladurner, 2003).

II-3 Les états de condensation de la chromatine

Une condensation progressive de la chromatine est particulièrement évidente lors de l'observation en microscopie photonique d'une division nucléaire révélant les chromosomes sous forme de filaments ou bâtonnets, leur raccourcissement atteignant un maximum en métaphase. Toutefois, la chromatine du noyau interphasique n'est aucunement homogène. Les examens en microscopie électronique ont permis de bien différencier deux types de régions : des régions de chromatine « diffuse » ou **euchromatine**, et des régions de chromatine « condensée » ou **hétérochromatine**, localisée principalement à la périphérie du noyau (**Figure 2**).

L'euchromatine est riche en gènes et très active sur le plan transcriptionnel. Quant à l'hétérochromatine, il est distingué: (1) l'**hétérochromatine constitutive (ou permanente)** pauvre en gènes, formée principalement de séquences répétées dont les plus connues sont celles des centromères et des télomères (Ridgway et Almouzni, 2001), (2) l'**hétérochromatine facultative** qui contient des séquences codantes pouvant adopter les caractéristiques structurales et fonctionnelles de l'hétérochromatine. Elle résulte de l'extinction de l'expression d'une fraction plus ou moins importante de l'euchromatine, comme dans le cas d'un chromosome X chez les femelles de Mammifères (Heard, 2004). Il a longtemps été supposé que l'hétérochromatine était transcriptionnellement inactive. On sait maintenant que certaines régions hétérochromatiniennes sont exprimées (Dimitri et al, 2005). Cependant, l'expression dans l'hétérochromatine des gènes qui y résident suppose manifestement une régulation adaptée. En effet, leur transfert par réarrangement dans l'euchromatine entraîne une répression de leur expression, ce qui constitue un "paradoxe" intéressant.

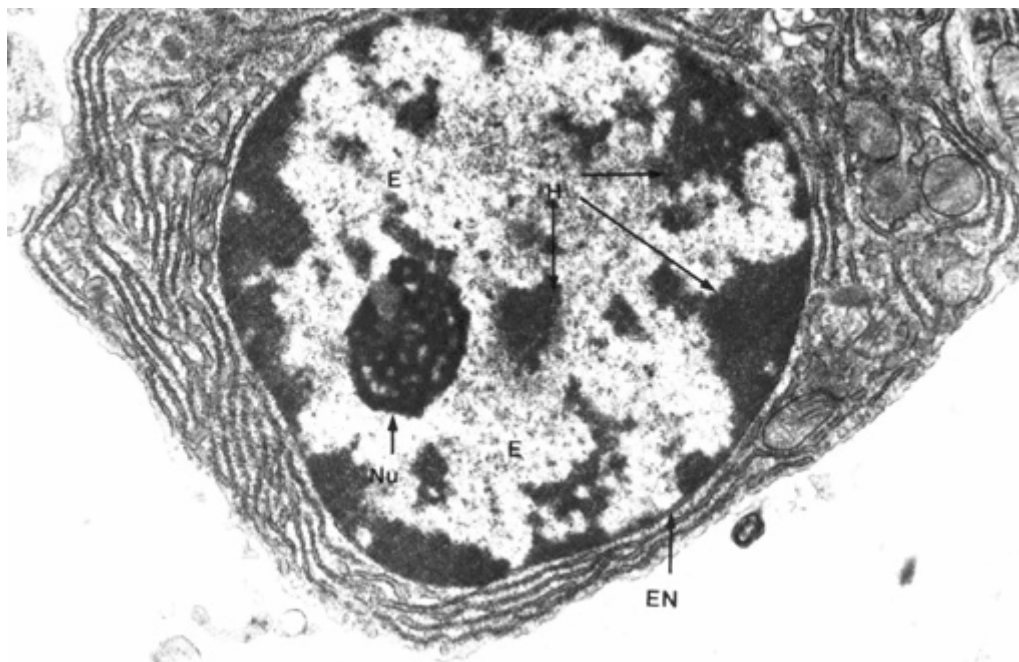


Figure 2: Aspect de la chromatine au sein d'un noyau observé sur coupe ultra-mince en microscopie électronique à transmission (d'après **Wheater, Burkitt et Daniels**). L'hétérochromatine (H), très opaque aux électrons, forme des blocs préférentiellement localisés à la périphérie tandis que l'euchromatine (E) est représentée par des régions diffuses plus internes. EN, enveloppe nucléaire ; Nu, nucléole. Grossissement : 15400.

La chromatine présente des niveaux supérieurs d'organisation schématiquement représentés dans la **figure 3**. En effet, lors de la mitose, l'ADN traverse des étapes de condensation et de ségrégation, nécessaires à la transmission d'une copie du génome à chacune des deux cellules-filles. Les changements drastiques qu'elles entraînent dans l'organisation de la chromatine sont le résultat de l'action concertée de nombreux effecteurs et régulateurs mitotiques.

Parmi les facteurs protéiques impliqués dans la condensation des chromosomes, on dénombre deux complexes multiprotéiques: la condensine et la cohésine, jouant respectivement un rôle dans la condensation de l'ADN et dans la cohésion inter-chromatides. Toutes deux sont constituées de protéines SMC et non-SMC (Jessberger, 2002).

Les protéines SMC («structural maintenance of chromosomes»), au nombre de six chez les Eucaryotes supérieurs (SMC-1, -2, -3, -4, -5 et -6), sont des protéines d'assez haut poids moléculaire (entre 1000 et 1400 acides aminés). Leur structure primaire est très conservée des bactéries à l'Homme. Elles s'hétérodimérisent pour former des complexes en

forme de V (Melby et al, 1998), aux extrémités desquels se situent des domaines ATPasiques, indispensables à leur activité (Löwe et al, 2001).

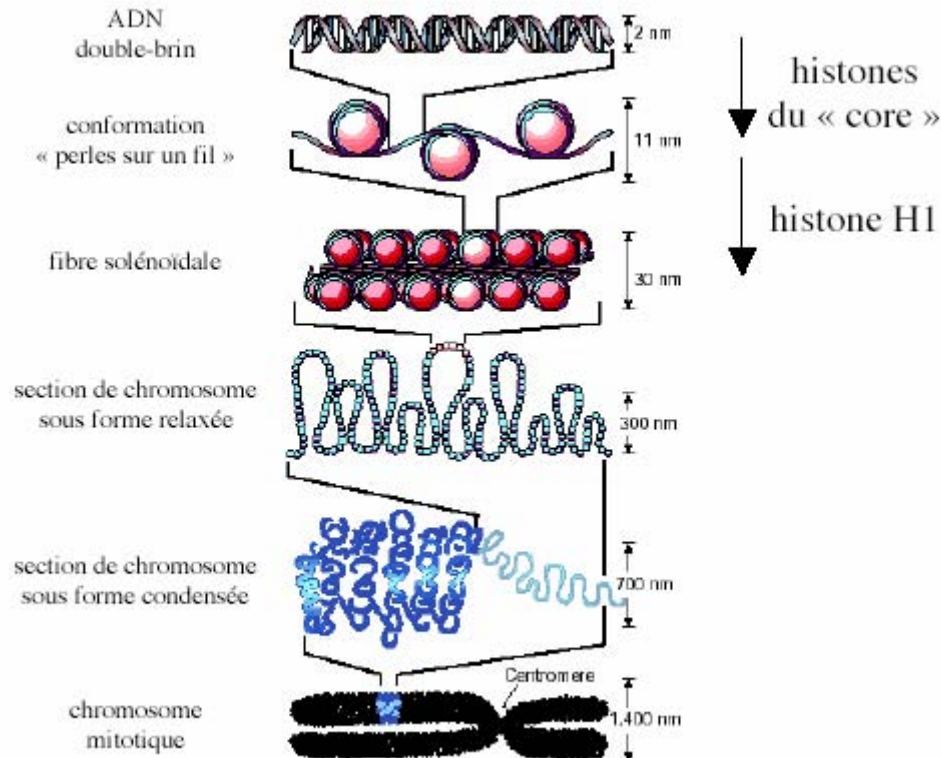


Figure 3: Les différents niveaux d'organisation de la chromatine. (D'après *Felsenfeld et Groudine, 2003*).

II-3-1 Le complexe condensine

Le complexe condensine (également appelé «complexe 13S» d'après son coefficient de sédimentation) résulte de l'association d'un dimère SMC-2/SMC-4 (CAP-E/CAP-C, ou «complexe 8S») et d'un trimère non-SMC CAP-D2/CAP-G/CAP-H (associées en «complexe 11S») (**Figure 4Aa**) (Hirano et al, 1997). Chacune de ces cinq sous-unités est essentielle chez la levure (Sutani et al, 1999). Sa fonction dans la condensation a été découverte à la suite de tests effectués dans des extraits d'ovocyte de *Xenopus laevis*, système couramment utilisé pour l'étude de la condensation des chromosomes. D'après des expériences d'immunodéplétion de ces extraits, la condensine est absolument nécessaire à la condensation des chromosomes (Hirano et al, 1997).

Les mécanismes par lesquels la condensine participe à cette condensation demeurent obscurs. Il semblerait qu'elle favorise, dans l'ADN auquel elle se lie, la création et la maintenance de boucles d'orientation positive (Kimura et Hirano, 1997), permettant ainsi un pontage intra-moléculaire (**Figure 4Ab**).

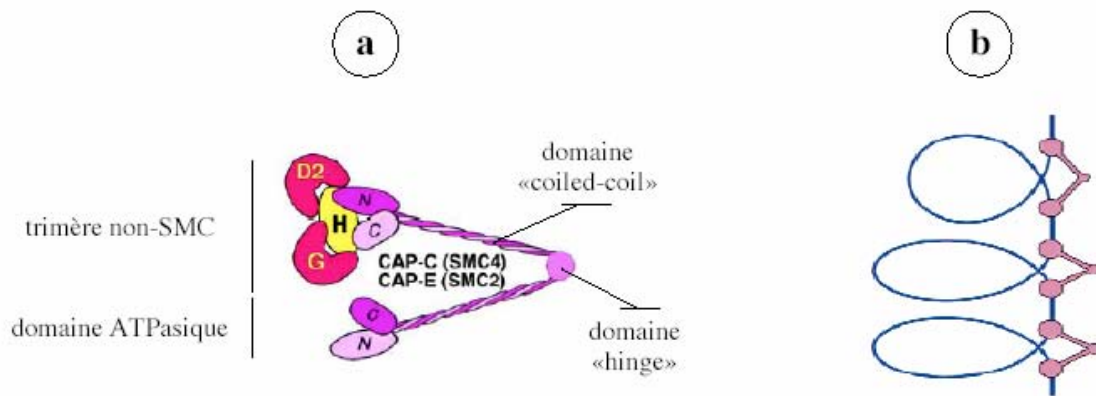


Figure 4A: Structure et modèle d'action de la condensine. a) La condensine est constituée d'un dimère SMC- 2/SMC-4 et d'un trimère de protéines non-SMC (d'après **Hirano, 2002**). b) Rôle proposé dans les pontages au sein d'une même molécule d'ADN avec formation de boucles (d'après **Hirano, 1999**).

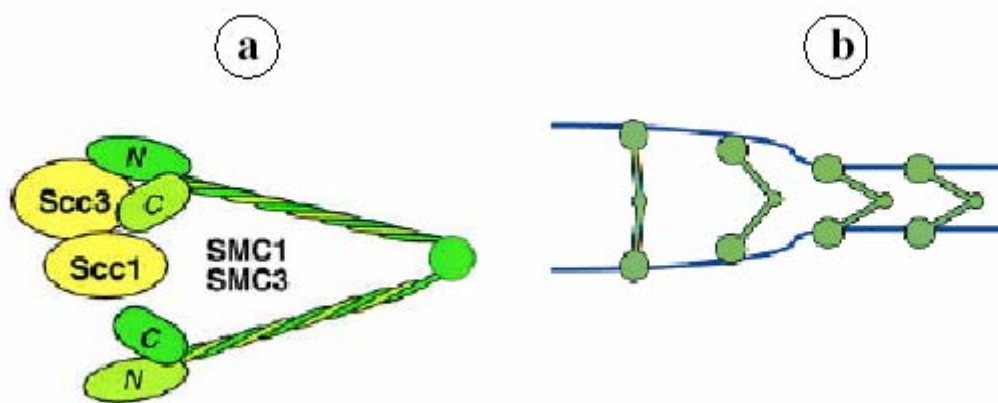


Figure 4B: Structure et modèle d'action de la cohésine. a) La cohésine est constituée d'un dimère SMC-1/SMC-3 et d'un dimère non-SMC (d'après **Hirano, 2002**). b) Rôle possible dans les pontages inter-moléculaires des ADN de deux chromatides (d'après **Hirano 1999**).

II-3-2 Le complexe cohésine

Le complexe cohésine est formé de l'hétérodimère SMC-1/SMC-3 ainsi que de deux protéines non-SMC, Scc1 et Scc3 (**Figure 4B**) (Losada et al, 1998). Des souches mutantes de levure pour la cohésine perdent des chromosomes à des fréquences anormalement élevées et présentent une séparation prématurée des chromatides (Michaelis et al, 1997).

Bien que les bases moléculaires de ses propriétés mécaniques n'aient pas encore été découvertes, il semble très probable que le complexe cohésine agisse en tant que "linker" inter-moléculaire entre deux molécules d'ADN, ce qui expliquerait sa capacité d'induire l'agrégation de l'ADN *in vitro* (Losada et Hirano, 2001b).

III- Le chromosome : éléments généraux de structure et de fonction

Le chromosome est un complexe nucléoprotéique caractérisé d'une part par l'information qu'il contient (les gènes), et d'autre part, par les éléments qui garantissent sa transmission de génération en génération. Cette transmission de l'information passe par une réplication semi-conservative et une ségrégation équitable des copies chromosomiques entre cellules-filles. Cela nécessite au moins trois types de composants : des origines de réplication, un centromère et deux télomères.

III-1 Origines de réplication

Pendant la phase S du cycle cellulaire eucaryote, la réplication de l'ADN de chaque chromosome démarre au niveau de nombreux sites, appelés « origines de réplication » (ORI) et espacés plus ou moins régulièrement le long du chromosome.

Les connaissances actuelles sur la nature de ces origines de réplication viennent essentiellement des levures, et particulièrement de *S. cerevisiae*. Le nombre total des ORI réparties sur les 16 chromosomes de *S. cerevisiae* est évalué entre 250 et 400 (Diffley, 1996). Ces ORI sont courtes (100 pb environ) et caractérisées par une séquence consensus de 17 pb appelée **ARS** (« Autonomously Replicating Sequence »). Cette séquence sert de site de reconnaissance au complexe multiprotéique **ORC** (« Origin Recognition Complex »), qui va servir à son tour de site d'assemblage du complexe de réplication (Gilbert DM, 2001). L'analyse comparative de diverses ARS (par usage de plasmides réplcatifs chez *S. cerevisiae*) révèle un consensus riche en thymine et nommé **ACS** (« ARS Consensus

Sequence ») (Diffley et stillman, 1990). Il est nécessaire aussi bien à l'activité ARS qu'à l'activité ORI (Desphande et Newlon, 1992).

La séquence *ARS1* montre une organisation reflétant les différentes interactions protéiques (Diffley et Cocker, 1992) (**Figure 5**). Elle comprend :

- une **boîte A**, contenant l'ACS flanquée de séquences variables mais fondamentales pour la fonction répliquative. Elle est reconnue par l'initiateur potentiel, c'est-à-dire le complexe ORC.

- trois **boîtes B** (B1, B2, B3), situées en 3' de la boîte A. B1 participe à la fixation de ORC, avec la boîte A (Rao et Stillman, 1995). B2 interagit avec une protéine de liaison à l'ADN simple brin (RP-A). Elle pourrait correspondre à un élément **DUE** (« DNA Unwinding Element ») que l'on retrouve sur des ARS dépourvus de boîte B2 et qui facilite l'écartement des deux brins parentaux d'ADN et l'intervention des enzymes de réplication (Natale et al, 1992). Enfin, la boîte B3 fixe le facteur transcriptionnel Abf1 (« ARS binding factor 1 ») qui intervient dans divers processus nucléaires, notamment la réplication de l'ADN (Diffley et Cocker, 1992).

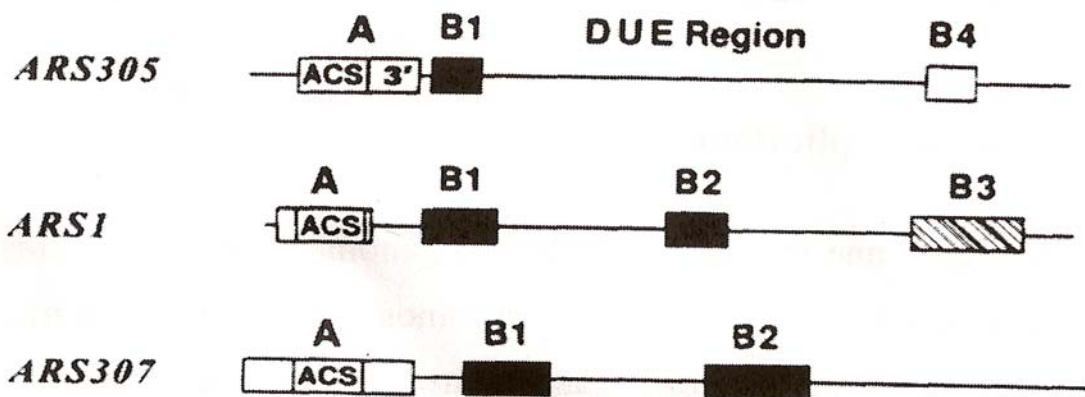


Figure 5 : Exemples de séquences ARS présentant une activité d'origine de réplication chez *Saccharomyces cerevisiae*.

Ces ARS ont toutes en commun la boîte A, contenant la séquence consensus ACS, et la boîte B1. La boîte B2 est remplacée par la région DUE dans *ARS305*. La boîte B3 de l'*ARS1* n'est pas conservée dans les deux autres ARS (Rao et Stillman, 1995).

Néanmoins, cette description pour *ARS1* est difficilement généralisable à d'autres ARS de *S. cerevisiae* (**Figure 5**).

Les origines de réplication ont été plus difficiles à identifier chez les autres Eucaryotes et des séquences spécifiques de liaison, telles que ORC, n'ont pas pu être identifiées (Schwob, 2004). Ainsi, les ORI de la levure *Schizosaccharomyces pombe* sont plus proches de celles décrites chez des métazoaires. Elles sont plus longues (1 kb environ) et ne présentent aucune séquence consensus, exceptée une courte séquence riche en adénine. Cependant, les seules origines caractérisées chez cette levure sont d'une efficacité faible avec, par exemple, des taux de 40 et 30 % pour ORI19 et ORI22 pendant la phase S. Par ailleurs, il a été montré que l'activation des origines de réplication ne se fait pas de façon coordonnée (revue : Patel et al, 2006). L'analyse de deux ARS (*ARS18* et *ARS68*) chez *Yarrowia lipolytica*, montre une nette divergence de séquence, par rapport à celles chez *S. cerevisiae*, *K. lactis* ou *S. pombe* (Fournier et al, 1993).

En conclusion, bien que des progrès aient été accomplis pour comprendre les mécanismes d'activation des différentes origines de réplication, la façon dont cette activation est coordonnée dans le temps et dans l'espace, reste peu claire (Schwob, 2004).

III-2 Centromères

Le centromère joue un rôle essentiel dans la ségrégation normale des chromosomes pendant la division cellulaire, en dirigeant la formation du kinétochore, complexe nucléoprotéique responsable de l'attachement au fuseau mitotique (Pluta et al, 1995). En dehors de quelques cas particuliers naturels (chromosomes mitotiques holocentriques de *Caenorhabditis elegans*) ou accidentels et délétères (chromosomes acentriques ou dicentriques), le chromosome est monocentrique c'est-à-dire avec un seul kinétochore par chromatide. Deux types de centromères ont été distingués en relation avec deux types de chromosomes eucaryotes : les centromères « ponctuels » d'environ 200 pb caractérisés initialement chez *S. cerevisiae*, et les centromères « régionaux » (40 kb au moins) rencontrés chez *S. pombe* et les métazoaires.

III-2-1 Centromères ponctuels

Ce sont les premiers centromères identifiés, fonctionnellement et moléculairement, au travers des nombreuses études portant sur le génome de *S. cerevisiae*. Ces séquences confèrent une stabilité et une ségrégation classique aux « minichromosomes » circulaires de *S. cerevisiae*, à raison d'une ou deux copies par cellule (Clarke et Carbon, 1980). Elles sont très courtes (près de 220 pb) et présentent une structure chromatiniennne compacte qui résiste à

l'action des nucléases (Bloom et Carbon, 1982) (**Figure 6A**). Les parties dites « minimales », directement impliquées, ont été déterminées pour chacun des 16 chromosomes et ont servi spécifiquement à la construction des chromosomes artificiels, dits **YAC** («Yeast Artificial Chromosomes»). S'étendant sur 125 pb environ, elles ne comportent aucune séquence répétée et ne sont pas transcrites. Elles sont peu différentes d'un chromosome à un autre et restent fonctionnellement interchangeables dans l'une ou l'autre orientation (Clarke, 1990). Trois blocs **CDE** (« Centromere DNA Element ») sont distingués (**Figure 6B**):

- **CDEI**, une courte séquence de 8 pb relativement conservée entre les chromosomes.
- **CDEII**, long de 78 à 86 pb et caractérisé par une séquence riche en AT (> 90 %). Cette séquence peut être formée aléatoirement avec les nucléotides A et T sans perte de fonction. Seule sa longueur doit être maintenue.
- **CDEIII**, d'une taille de 26 pb et présentant un motif central CCG essentiel au fonctionnement du centromère.

Ces blocs sont souvent flanqués de régions transcrites. Par exemple, dans le cas du chromosome XI, la transcription du gène *spo15* s'arrête environ 300 pb en amont de CDEI et celle du gène *met14* commence environ 500 pb en aval de CDEIII (**Figure 6A**). Les trois blocs procurent à la molécule d'ADN une forme incurvée (Bechert et al, 1999), visible en microscopie à force atomique (**Figure 7**) lorsque l'un des éléments protéiques du kinétochore (protéine CBF3) est lié au bloc CDEIII (Pietrasanta et al, 1999).

Le modèle de centromère ponctuel est retrouvé au moins en partie chez d'autres levures bourgeonnantes. Chez *Kluyveromyces lactis*, cinq des six centromères ont pu être clonés et ont permis de définir trois CDE par centromère, par analogie avec ceux de *S. cerevisiae* (Heus et al, 1993). Leurs séquences consensus (**Figure 6B**) montrent une très grande conservation entre les genres *Kluyveromyces* et *Saccharomyces*. Notons que l'orientation de CDEI est variable d'un chromosome à l'autre chez *K. lactis* alors qu'elle est conservée chez *S. cerevisiae*. La différence majeure est observée au niveau du bloc CDEII qui est deux fois plus long chez *K. lactis* (160 pb). Ces centromères ne sont pas interchangeables d'un genre à l'autre (Heus et al, 1994). *Candida glabrata* possède des séquences centromériques d'une longueur de 153 pb et trois CDE ont aussi été mis en évidence (Kitada et al, 1997).

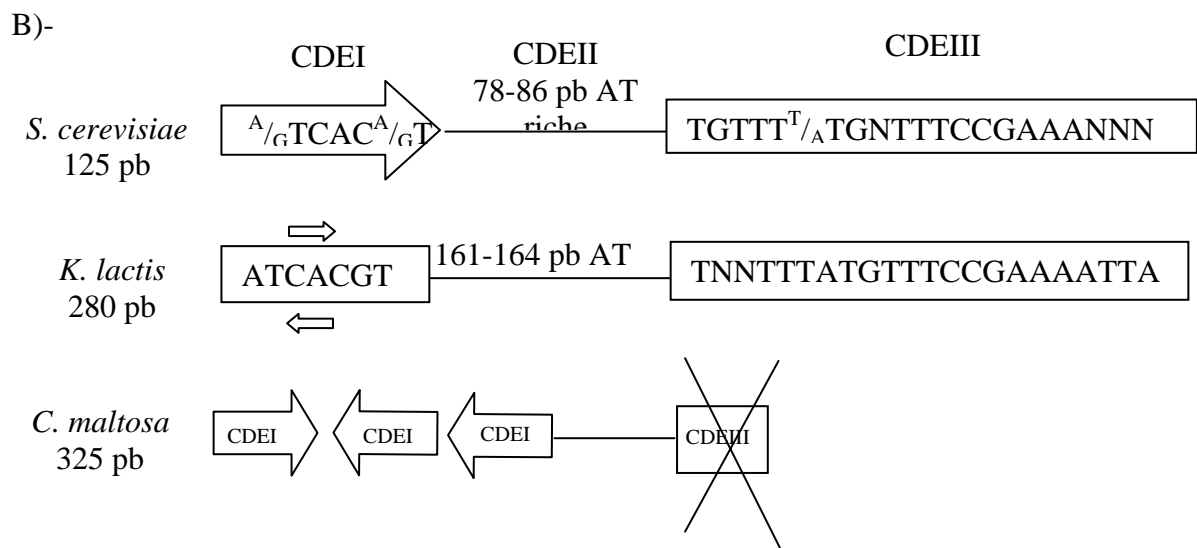
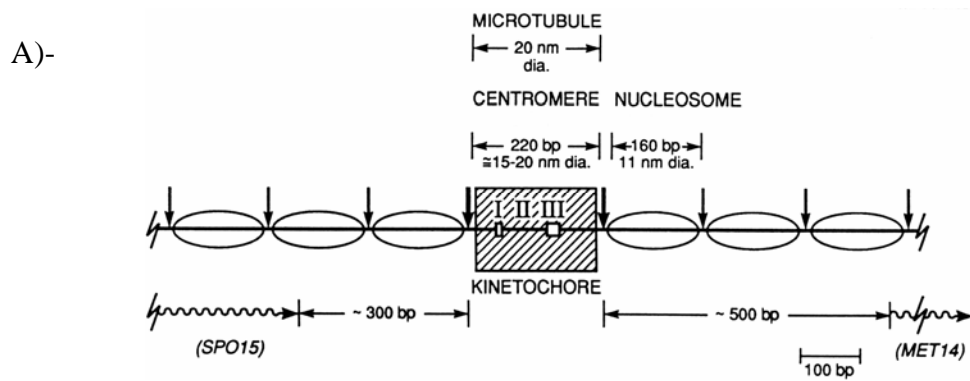


Figure 6 : Exemples de centromères ponctuels.

A/ Structure du centromère du chromosome XI de *Saccharomyces cerevisiae* avec les trois boîtes caractéristiques CDEI, CDEII et CDEIII. Cette région est comprise dans 220 pb, au milieu d'une structure nucléosomale classique, présentant des gènes transcrits (*SPO15* et *MET14* dans cet exemple) (Clarke, 1990).

B/ Boîtes CDEI, CDEII et CDEIII des centromères ponctuels chez trois levures. La taille totale de la région impliquée est indiquée. Différentes orientations des boîtes CDEI sont indiquées par des flèches. Chez *Candida maltosa*, la séquence de CDEI est ATCAATGA (adapté d'après Clarke, 1990 ; Heus et al, 1993 ; Ohkuma et al, 1995).

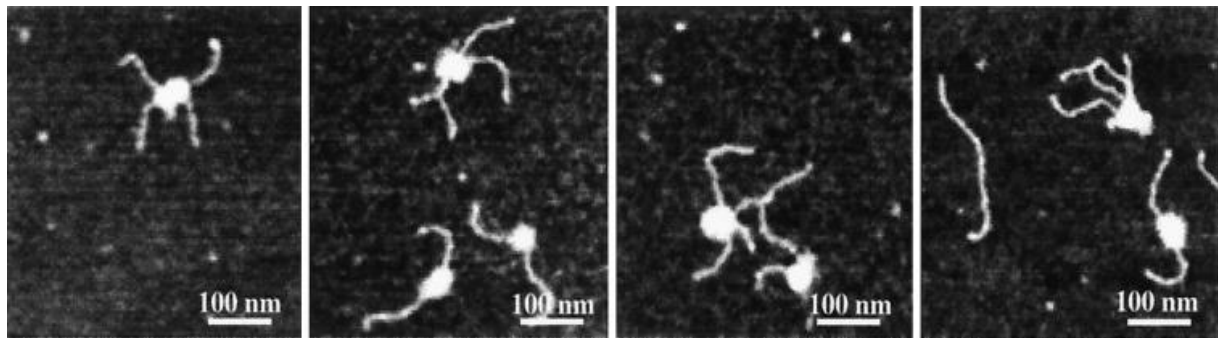


Figure 7: Observation en microscopie à force atomique des courbures de l'ADN suite à la fixation de CBF3 à CDEIII (Pietrasanta et al, 1999).

III-2-2 Centromères régionaux

Les centromères régionaux sont plus complexes. Chez *S. pombe*, la taille de la région impliquée dans la fonction centromérique varie de 40 à 100 kb selon les souches et les chromosomes (Steiner et al, 1993). Cette taille peut atteindre 5 Mb pour des centromères humains (Floridia et al, 2000).

Un centromère régional est formé en grande partie d'éléments répétés non transcrits. Chez *S. pombe*, une région centrale longue de 4 à 7 kb est encadrée par des éléments répétés disposés symétriquement (Clarke et Baum, 1990). Les répétitions dites **K**, **L** et **M**, d'après la nomenclature de Wiens et Sorger (1998) forment des blocs de 5 kb environ. Un segment de moins de 20 kb, couvrant le domaine central et les répétitions de type K, peut être suffisant pour obtenir un centromère fonctionnel (Baum et al, 1994). La **figure 8** illustre l'organisation de trois centromères de la souche Sp223. L'originalité des centromères régionaux réside aussi dans leur mode d'activation. Deux séquences centromériques peuvent cohabiter sur un même chromosome mais une seule est active, contrairement au modèle précédent. Cette activation est sous le contrôle de facteurs épigénétiques et dépend donc de l'organisation chromatinienne (Wiens et Sorger, 1998).

Le séquençage des chromosomes 2 et 3 de *Plasmodium falciparum* a permis de mettre en évidence une séquence de 2,6 kb extrêmement riche en AT qui présente des homologies avec les centromères de *S. pombe* (Bowman et al, 1999). Ces centromères putatifs sont cependant beaucoup plus courts et pourraient être les représentants d'une troisième classe de centromères.

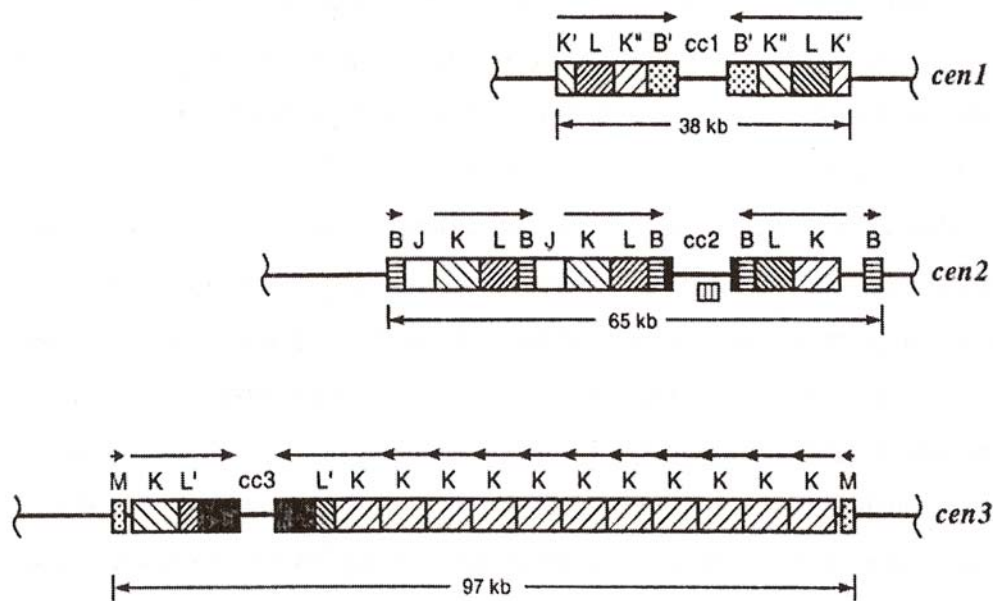


Figure 8 : Organisation schématique des centromères régionaux des trois chromosomes de *Schizosaccharomyces pombe*, souche Sp223 (Steiner et al, 1993).

Les différents blocs de répétitions sont indiqués par des boîtes (B et dérivé B'; J; K et dérivés K' et K''; L et dérivé L'; M) et encadrent la séquence du cœur du centromère (cc, « central core »).

III-3 Extrémités chromosomiques

Le concept de télomère est né en 1938, suite aux études de H.J. Müller sur les chromosomes de la drosophile. Celui-ci observa que des délétions ou des inversions induites par irradiation aux rayons X de l'ensemble du génome étaient quasiment indétectables dans les régions terminales des chromosomes. Il en conclut que les extrémités des chromosomes devaient posséder une structure particulière capable de stabiliser les chromosomes. À la même époque, B. McClintock démontra que des chromosomes du maïs ayant subi des cassures double brin étaient capables de fusionner entre eux, contrairement à leurs extrémités qui demeuraient stables. L'organisation des télomères a pu être caractérisée chez différentes espèces et la découverte de l'activité télomérase a été cruciale pour la compréhension de la réplication des extrémités des molécules d'ADN linéaire. Les cœurs chromosomiques sont séparés des télomères par des régions particulièrement dynamiques, les subtélomères, dont l'importance dans les processus de recombinaison sera ici soulignée.

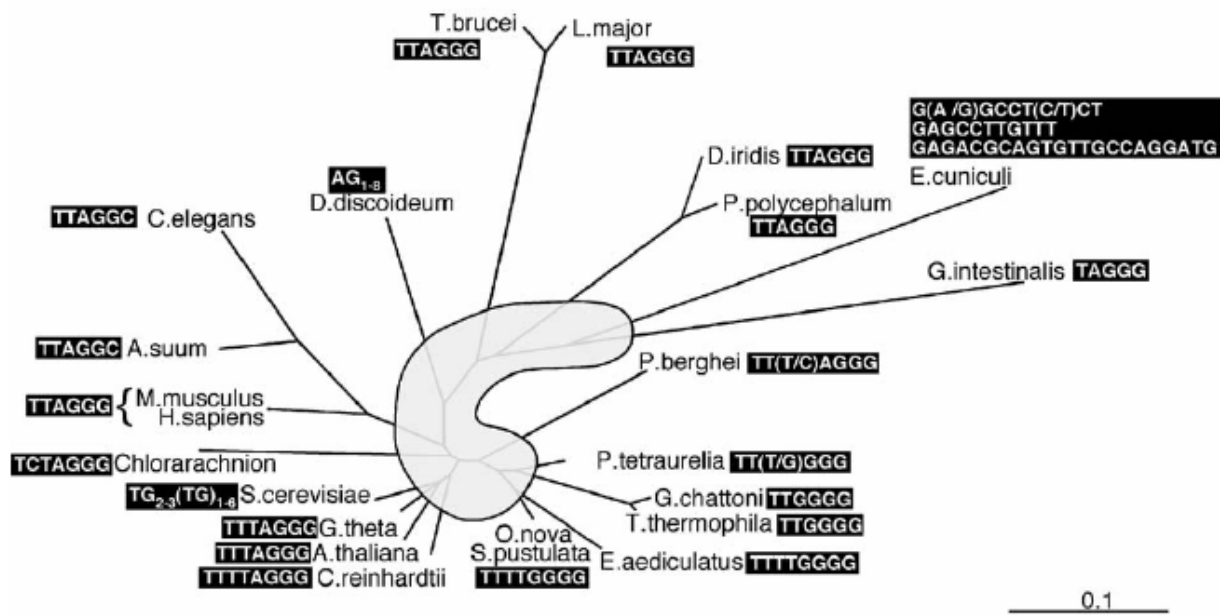
III-3-1 Télomères

Les télomères sont des structures nucléoprotéiques complexes qui protègent les extrémités des chromosomes linéaires chez les Eucaryotes (Blackburn, 2001) ainsi que chez certaines Eubactéries (Kobryn et Chaconas, 2001).

Les séquences d'ADN télomérique peuvent s'étendre sur plusieurs kilobases chez l'Homme (~10 kb) et certains protozoaires parasites (Scherf et al, 2001), ou sur seulement quelques centaines de bases chez les levures (~300 pb chez *S. cerevisiae*) et les Ciliés. De manière générale, il s'agit de courtes séquences répétées, identiques sur tous les chromosomes d'un même génome. Le motif répété est toujours riche en guanine (**Figure 9**) : TTAGGG chez les mammifères et la plupart des Eucaryotes (Shore, 2001), TTGGGG chez le Cilié *Tetrahymena thermophila* (Blackburn et al, 1978) ou (TG)₁₋₆TG₂₋₃ chez les levures comme *S. cerevisiae* (Cohn et al, 1998). Le nombre de répétitions peut varier non seulement d'une espèce à l'autre mais aussi entre les individus d'une même espèce et entre les différents chromosomes d'une même cellule. En effet, les régions télomériques sont fréquemment soumises à des phénomènes d'élongation et de raccourcissement (Barry et al, 2003). Chez tous les organismes étudiés, le brin riche en guanine (G-riche) correspondant à l'extrémité 3', est plus long que le brin complémentaire riche en cytosine (C-riche), créant ainsi une extension simple brin en 3' du chromosome (**Figure 10A**). Cette extension simple brin joue un rôle essentiel dans la protection de l'extrémité chromosomique (de Lange, 2001).

L'ADN télomérique est associé directement ou indirectement à un ensemble de protéines, formant une structure complexe, appelée **télosome (Figure 10B)**. Ces protéines reconnaissent des séquences d'ADN spécifiques et sont nécessaires à la réplication et au maintien de l'intégrité des télomères.

A



B

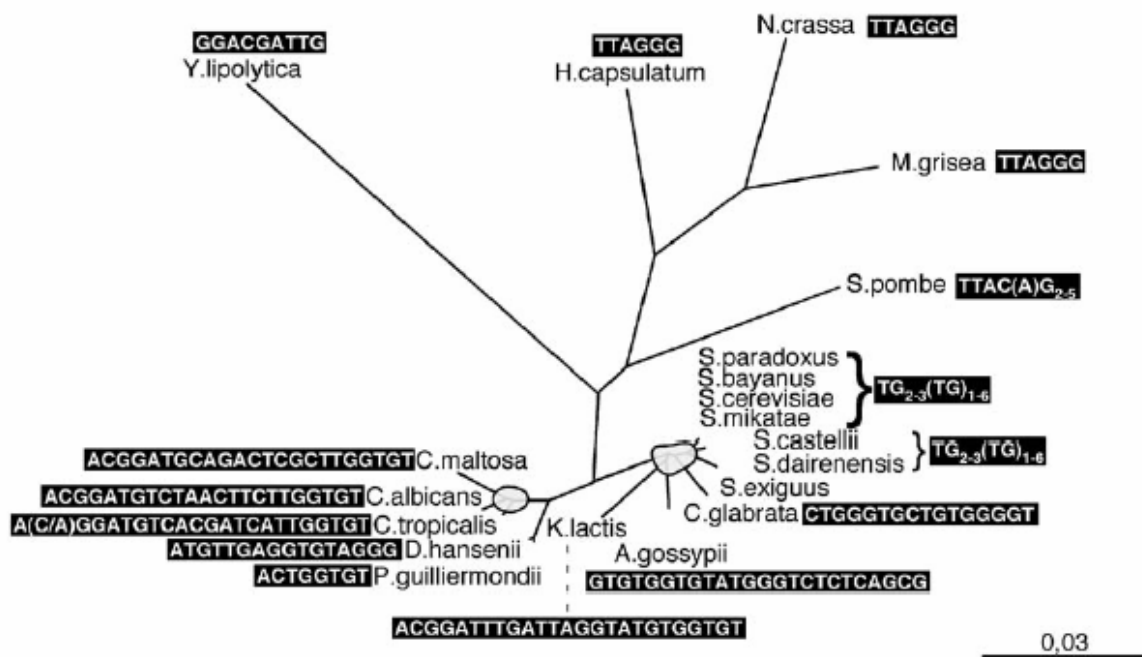
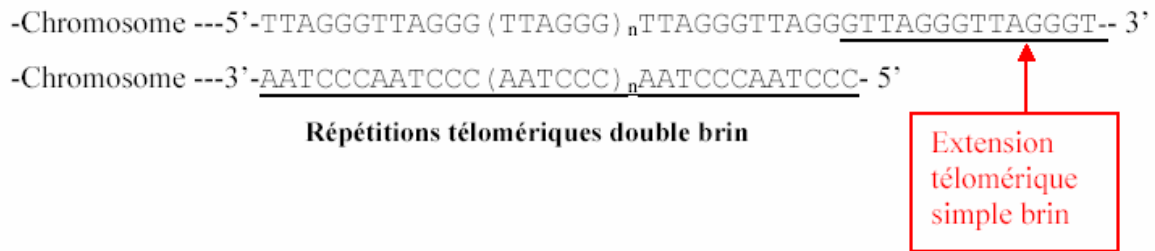


Figure 9 : Répétitions télomériques.

A / Chez des représentants de divers eucaryotes. **B/** Chez les Champignons Ascomycètes. Les arbres phylogéniques sont basés sur la comparaison des séquences d'ARNr 18S. (Teixeira et Gilson, 2005)

A



B

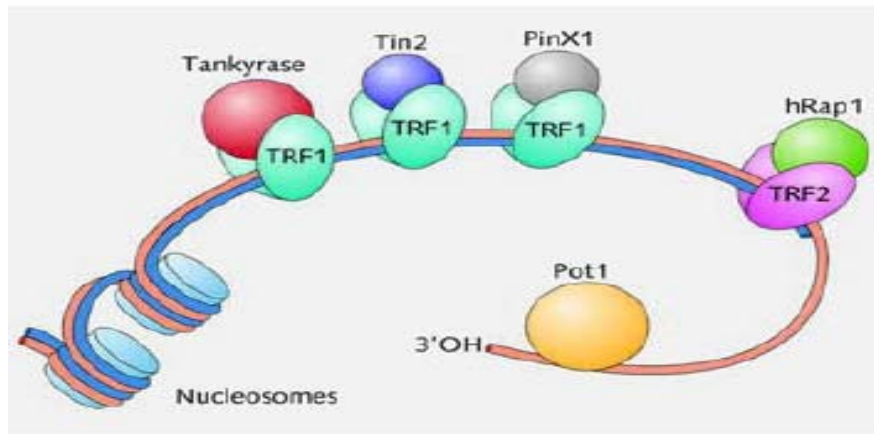


Figure 10: Télomère humain. **A** / Structure primaire de l'ADN d'un télomère. **B**/ Structure du télosome. Les rôles des différentes protéines du complexe dans la protection et la régulation de la longueur des télomères sont commentés dans le texte. (Rhodes et al, 2003)

III-3-1-1 Régulation de la longueur des télomères

Rôle de la télomérase:

La machinerie classique de réplication de l'ADN étant incapable de copier une extrémité libre (Henson et al, 2002), la réplication des télomères fait appel à une enzyme particulière, la télomérase. En l'absence de cette enzyme, les extrémités des chromosomes subiraient systématiquement le raccourcissement, comme illustré en **Figure 11** (Rieko Ohki et al, 2001).

1. La réplication des 2 brins d'ADN est réalisée dans des directions opposées à partir de l'origine de réplication.
2. A la fin de la réplication, seule une amorce d'ARN reste présente au niveau de l'extrémité de chaque brin néoformé.
3. Les dernières amorces sont éliminées par l'action d'une 5'→3' exonucléase laissant une extrémité simple brin.
4. Chaque cycle de réplication génère un raccourcissement des extrémités télomériques.

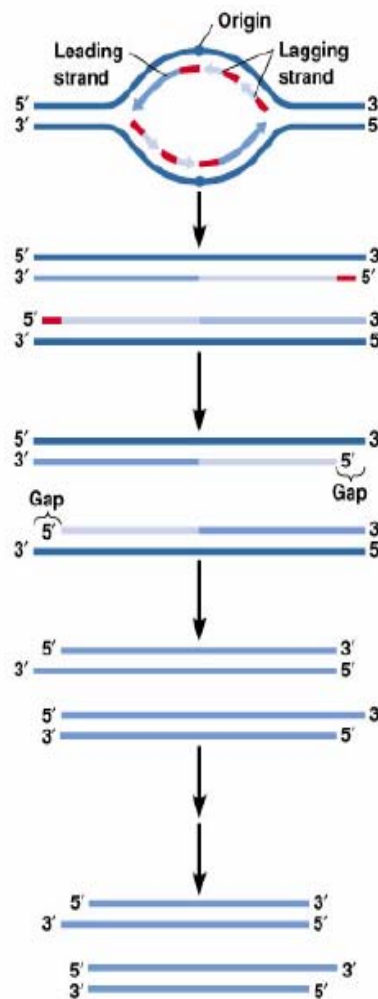


Figure 11: Problème posé par la réplication au niveau des extrémités chromosomiques. L'ADN polymérase polymérise les désoxyribonucléotides uniquement dans le sens 5'→3' et à partir de l'extrémité 3' libre d'une courte amorce ARN. L'élimination des amorces ARN aux extrémités du chromosome constitue alors une cause de raccourcissement du chromosome.

La télomérase (**Figures 12 A, B**) est une enzyme de nature ribonucléoprotéique (~1000 kDa chez l'Homme ; Beattie et al, 2001) qui comprend deux sous-unités essentielles à son activité (Bachand et al, 2001b) :

- la sous-unité catalytique TERT (« TELomerase Reverse Transcriptase »), protéine présentant une activité de transcriptase inverse et fortement conservée chez les organismes uni- et pluricellulaires (Lingner et al, 1997).

- la sous-unité TERC (« TELomerase RNA Component), également référencée sous les sigles TR ou TER (Feng et al, 1995), molécule d'ARN contenant une courte séquence matricielle pour la transcription en ADN télomérique.

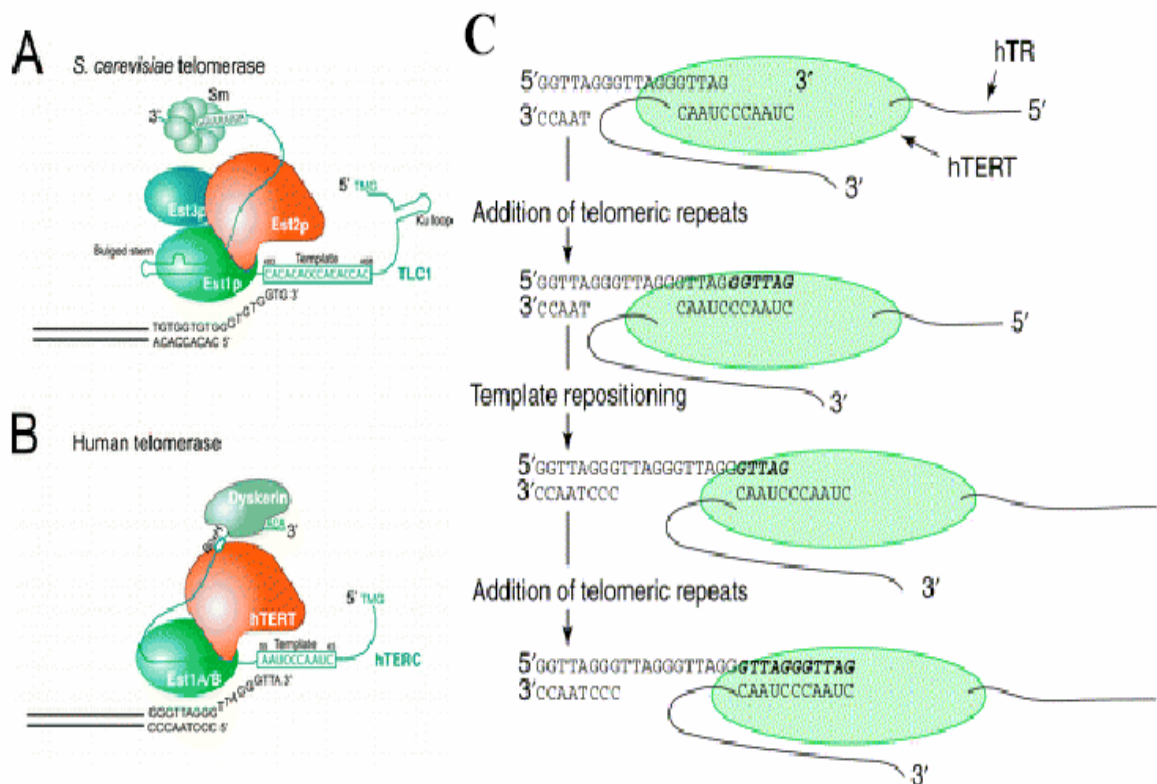


Figure 12: Représentation schématique des complexes télomériques chez *Saccharomyces cerevisiae* (A) et chez l'Homme (B). C/ Cycle d'addition des répétitions télomériques par la télomérase humaine. Fixée à l'extrémité du chromosome, cette transcriptase inverse ribonucléoprotéique ajoute un hexamère TTAGGG à l'extension simple brin de l'ADN. Une courte séquence de son ARN sert de matrice (« template »). Après synthèse de l'hexamère, l'enzyme se repositionne pour permettre un nouveau cycle d'élongation. (White L. K., et al, 2001).

Positionnée sur le télomère, l'enzyme fonctionne de manière processive en allongeant l'extension monocaténaire 3' par addition de nucléotides G et T dans l'ordre déterminé par le site matriciel de TERC (**Figure 12 C**). Le brin G-riche suffisamment long redevient accessible au complexe de réplication pour la synthèse du brin complémentaire.

Intervention d'autres protéines:

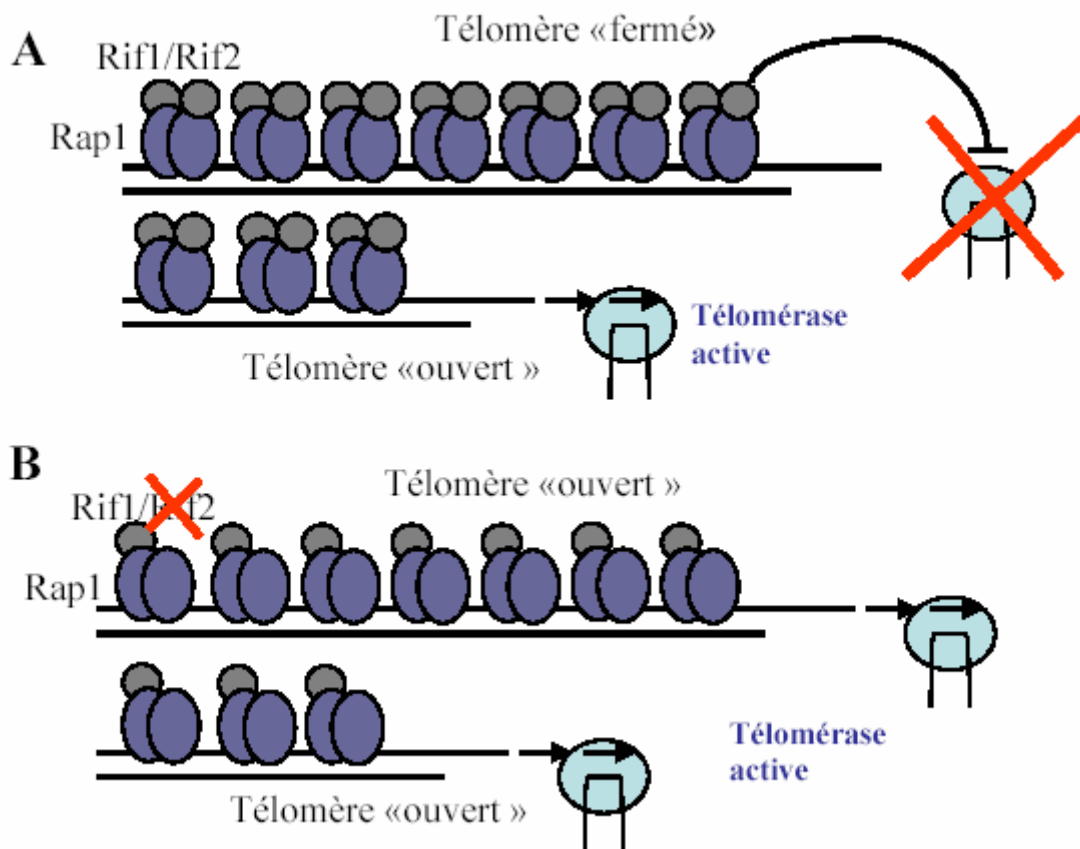


Figure 13 : Modèle du commutateur binaire entre l'état « ouvert » et « fermé » du télomère pour le contrôle de l'élongation des télomères. Rap1 se fixe au télomère et agit comme un régulateur négatif de l'activité télomérase d'une façon proportionnelle à son taux de fixation. Lorsque le télomère se raccourcit, l'inhibition induite par Rap1 est levée et le télomère peut passer à l'état « ouvert », accessible à la télomérase.

A/ En présence de Rif1 et de Rif2, seuls les télomères courts deviennent « ouverts » et seront allongés. **B/** En l'absence de Rif1 ou de Rif2, tous les télomères adoptent l'état « ouvert » et peuvent alors être allongés (d'après **Riou et al, 2005**).

Chez *S. cerevisiae*, la protéine **RAP1** se fixe sur le télomère, permettant de recruter d'autres facteurs (Rif1 et Rif2) pour former le complexe majeur de régulation de la longueur des télomères par un rétrocontrôle négatif de l'activité télomérasique (Hardy et al, 1992). Cette protéine possède deux domaines Myb de liaison à l'ADN. La télomérase allonge préférentiellement les télomères de petite taille qui sont sous un état dit « ouvert », les plus longs tendant à prendre l'état alternatif « fermé » inaccessible à l'enzyme. La régulation via RAP1 s'effectue de façon proportionnelle au nombre de molécules protéiques fixées (**Figure**

13). L'inhibition de la télomérase est d'autant plus forte qu'il y a plus de molécules RAP1 sur un long télomère. Quand le télomère se raccourcit, le nombre de molécules RAP1 diminue et le télomère devient alors plus accessible à la télomérase (Marcand et al, 1997). L'absence de Rif1 ou Rif2 renforce cette accessibilité (Wotton et Shore, 1997). D'autres études ont montré que la surexpression de RAP1 provoque un rallongement significatif des télomères (Li et al, 2000).

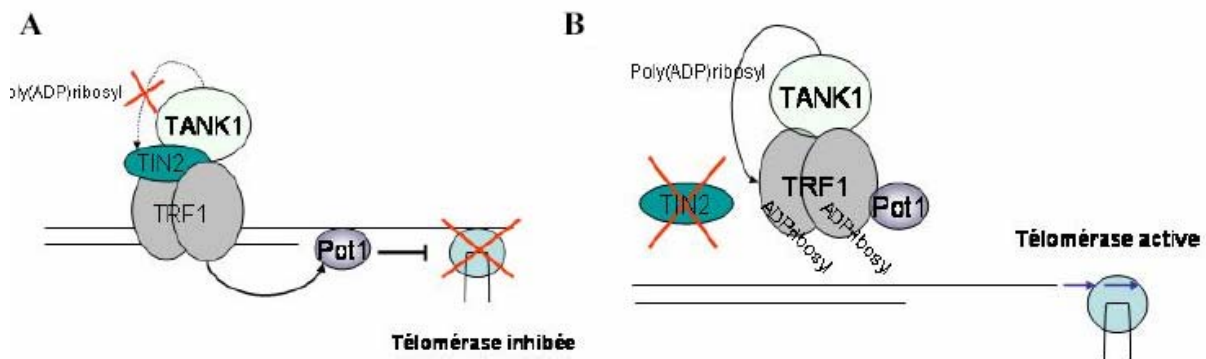


Figure 14: Le facteur TIN2 est un régulateur de l'activité de la tankyrase. **A/** TIN2 se fixe à TRF1 et le protège de l'inactivation par la tankyrase (TANK1). **B/** Lorsque TIN2 est inactivé, l'activité poly(ADP) ribosyltransférase de TANK1 diminue l'affinité de TRF1 pour l'ADN télomérique double brin, ce qui empêche la fixation de POT1 au simple brin télomérique. Le simple brin devient alors accessible à la télomérase (Riou et al, 2005).

Chez l'Homme, le rôle de RAP1 est joué par la protéine **TRF1** (« Telomeric Repeat binding Factor 1 »), un homodimère avec un domaine Myb-like (Steensel et de Lange 1997). Elle recrute aussi d'autres protéines pour former un complexe de régulation (Loayza et de Lange 2003). TRF1 agit bien comme un régulateur négatif de la télomérase. La surexpression de TRF1 entraîne une diminution de la taille des télomères alors que la surexpression d'un dominant négatif de TRF1 entraîne une élongation des télomères (Smogorzewska et al, 2000). Plusieurs protéines partenaires de TRF1 ont été identifiées et sont capables d'influencer cette régulation négative. C'est le cas de la tankyrase 1 (TANK1), une protéine de la famille des poly(ADP-ribose) polymérase. La polyADP-ribosylation de TRF1 provoque l'inhibition de la fixation de ce facteur aux télomères (Smith et al, 1998). Dans des cellules humaines ayant une activité télomérasique, la surexpression de TANK1 induit bien l'élongation progressive des télomères (Smith et de Lange, 2000). L'activité de la tankyrase sur TRF1 est elle-même

régulée par le facteur TIN2 (**Figure 14**) (Ye et de Lange, 2004a). TIN2 contribue à l'accumulation de TRF1 au niveau des télomères. L'inactivation de TIN2 autorise la polyADP-ribosylation de TRF1 par TANK1, ce qui empêche indirectement la fixation d'une protéine protectrice (POT1) du simple brin télomérique qui devient alors plus accessible à la télomérase (Riou et al, 2005).

III-3-1-2 Protection contre les recombinaisons

La présence d'une extrémité d'ADN libre dans un génome constitue un « signal d'alarme » pour la cellule (Van Gent et al, 2001). La détection rapide de cette anomalie conduit soit à sa réparation, soit à un arrêt du cycle cellulaire et parfois à l'apoptose. L'architecture particulière des télomères assure une séquestration des extrémités chromosomiques qui permet aux cellules de distinguer ces extrémités des cassures accidentelles pouvant survenir au sein des chromosomes, en vue de les protéger des fusions chromosomiques, des recombinaisons, des exonucléases et de la réparation.

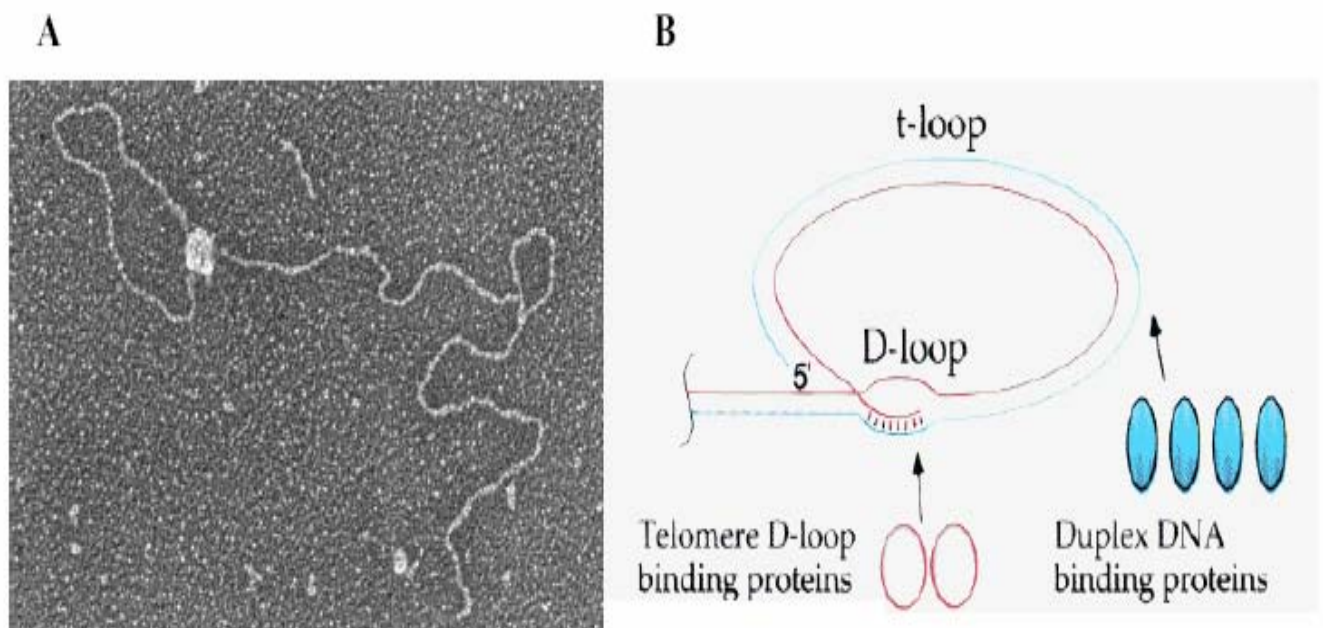


Figure 15: A/ Image en microscopie électronique d'une boucle t obtenue en présence de TRF2 (Griffith et al, 1999). B/ Modèle de formation d'une boucle t (« t-loop » pour « telomere loop ») (Greider, 1999). Sur l'image A, la petite boucle D (« D loop » pour « displacement loop »), résultant de l'insertion de l'extension simple brin, est masquée par la protéine de liaison spécifique.

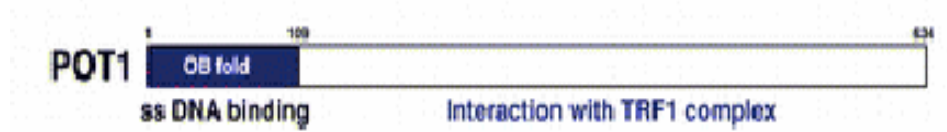
Des observations en microscopie électronique chez quelques organismes, tel le Cilié hypotriche *Oxytricha fallax*, ont montré que le télomère est replié sur lui-même, formant une « boucle t » (« t-loop »; Murti et Prescott 1999). Cette boucle résulte de l'insertion de l'extrémité 3' simple brin libre dans une portion double brin du télomère, comme illustré en **figure 15**.

Chez l'Homme, la protéine **TRF2** se liant à l'ADN double brin est essentielle à la protection des télomères en intervenant dans la formation des boucles t (van Steensel et al, 1998). Comme TRF1, elle présente un domaine Myb et fonctionne comme un homodimère. Pour la fonction de protection, TRF2 recrute hRAP1, orthologue de la protéine RAP1 de *S. cerevisiae*. Chez *S. pombe*, RAP1 ne possède pas de domaine de liaison à l'ADN et interagit avec la protéine Taz1p (« Telomere associated in *Schizosaccharomyces pombe* »), très proche de la TRF2 humaine tant au niveau structural que fonctionnel (de Lange 2002).

Une autre protéine **POT1** (« Protection Of Telomere 1 ») impliquée dans la protection des télomères, a initialement été identifiée chez le Cilié *Oxytricha nova* (Baumann et Cech, 2001). Elle se fixe à l'ADN simple brin. Cette protéine possède en effet à son extrémité N-terminale un domaine « Oligonucleotide Binding fold » (« OB-fold ») de fixation à l'ADN simple brin qui lui permet de reconnaître spécifiquement TAGGGTTAG comme séquence minimale de reconnaissance (**Figure 16**). Son extrémité C-terminale présente un domaine de reconnaissance à un complexe associé à TRF1. Des homologues ont été retrouvés chez l'Homme (hPOT1 ; Lei et al, 2004) et *S. cerevisiae* (CDC13 ; Theobald et Wuttke, 2004).

Le complexe **Ku** jouerait un rôle important dans les interactions des télomères avec la matrice nucléaire (Laroche et al, 1998) et dans la stabilisation des télomères. Chez les mammifères, le complexe Ku fait partie d'un complexe protéique, de la famille du gène *ATM*, qui a une activité kinase ADN-dépendante. Ce complexe est normalement nécessaire à la réparation de l'ADN et aux recombinaisons (Dynan et Yoo, 1998).

A



B

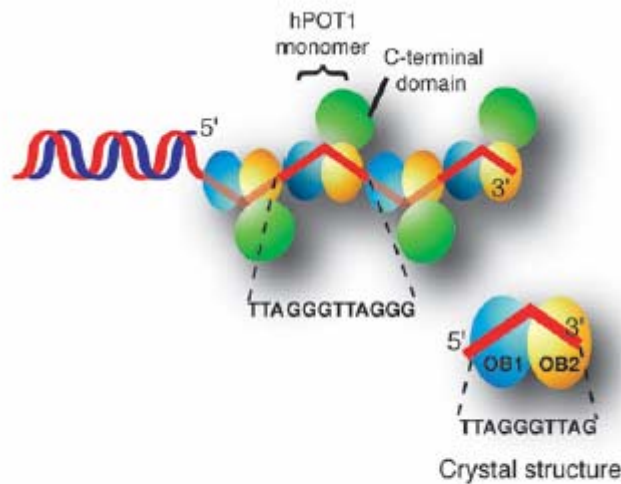


Figure 16: *A/ Localisation des domaines de liaison à l'ADN simple brin et à TRF1 dans la protéine de protection télomérique POT1. B/ Illustration schématique montrant comment l'homologue humain hPOT1 (avec deux domaines « OB-fold » : OB1 et OB2) se fixe à l'extrémité 3' simple brin des télomères (Lei et al, 2004).*

III-3-1-3 Regroupement des télomères dans le noyau et « effet de position »

Il est connu depuis longtemps que les télomères ont tendance à se positionner à la périphérie du noyau, une observation qui a été nettement confirmée par la microscopie à haute résolution (Klein et al, 1992). De plus, les télomères se regroupent en formant plusieurs petits foyers (« clustering »).

Chez *S. pombe*, l'attachement des télomères à la périphérie nucléaire s'effectue sous l'influence de la protéine Taz1. Il facilite la formation des paires de chromosomes homologues et par conséquent les recombinaisons par crossing-over durant la méiose (Cooper et al, 1998). Chez *S. cerevisiae*, le complexe protéique Ku est nécessaire au regroupement des télomères, Ku70 pouvant se lier indirectement aux pores nucléaires (Galy et al, 2000).

Le phénomène de regroupement a des conséquences notables sur l'expression des gènes et la dynamique de ces régions chromosomiques. En effet, il induit une inhibition de certains gènes subtelomériques (« effet de position aux télomères »; Gottschling et al, 1990) et de fréquentes recombinaisons entre chromosomes hétérologues ou recombinaison ectopique (Barry et al, 2003).

Il a été montré que l'introduction du gène *URA3* en position subtelomérique sur des chromosomes de *S. cerevisiae* ou *S. pombe* entraîne une répression transcriptionnelle réversible de ce gène (Sandell et al, 1994). Les premières données indiquaient que plus le gène inséré est proche du télomère et plus le télomère est long, plus l'effet de position est fort (Kyrion et al, 1993). Chez *S. cerevisiae*, une telle inhibition a été mise en relation avec la localisation périphérique des télomères (Feuerbach et al, 2002). Toutefois, il n'est pas encore clair s'il s'agit d'un mécanisme majeur de régulation de l'expression des gènes *in vivo*. Une étude comparative de gènes « naturellement » subtelomériques chez *S. cerevisiae* a suggéré que seulement un gène de ce groupe était soumis à l'effet de position aux télomères (Vega-Palas et al, 2000). Une autre analyse portant sur une plus grande partie du génome a fait conclure que seulement 16% des gènes à localisation subtelomérique peuvent subir cet effet (Wyrick et al, 1999).

Deux types de répression (« silencing ») ont été distingués en fonction de l'intervention ou non du complexe **Ku**. La répression Ku-dépendante est la plus forte. Outre sa participation à la régulation de la longueur des télomères, Ku intervient normalement dans la réparation d'ADN en se dissociant des subtelomères après détection d'une coupure d'ADN double brin dans une autre région du génome (Martin et al, 1999). La répression Ku-indépendante est plus encline à être levée. Les facteurs impliqués sont les protéines **Sir**, protéines inhibitrices qui se fixent aux régions subtelomériques par l'intermédiaire de Rap1 (**Figure 17**). Chez *Trypanosoma sp*, le complexe Ku est associé au maintien de la structure des télomères (longueur des télomères) mais il ne semble avoir aucune influence sur l'expression télomérique des gènes *vsg* (Conway et al, 2002).

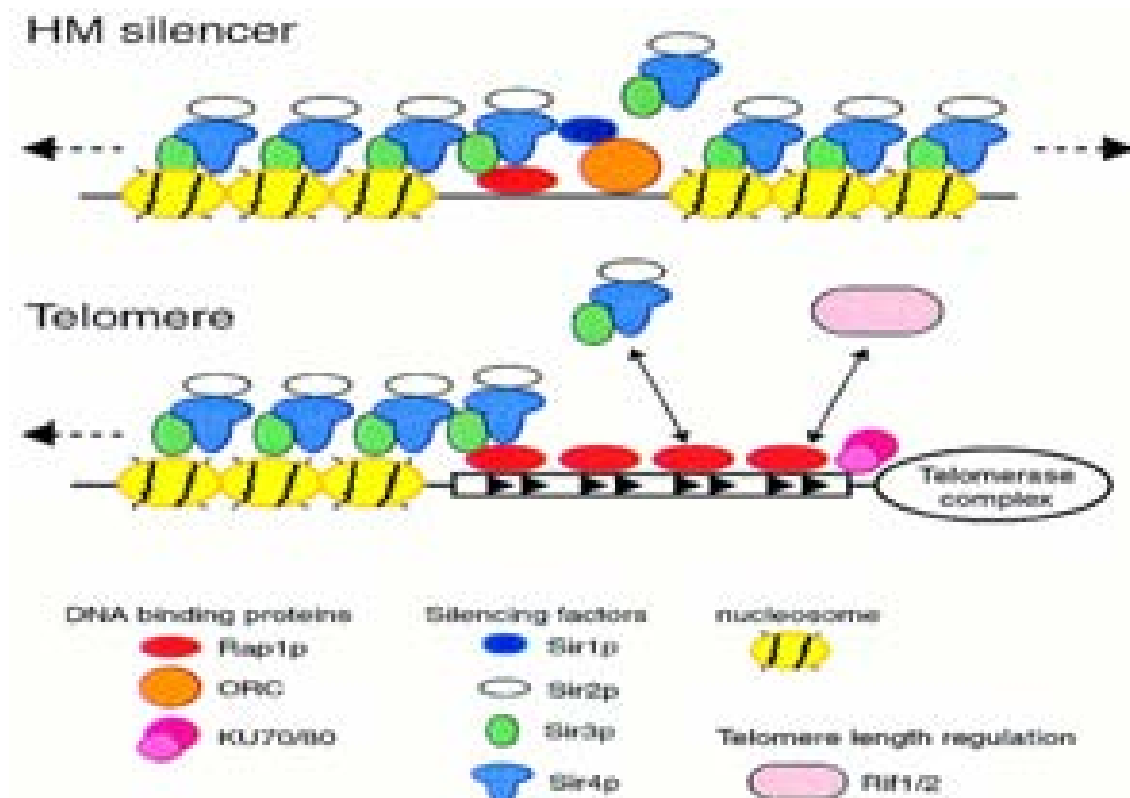


Figure 17: Recrutement de la protéine Rap1 (ici représentée par un ovale rouge) pour débloquent l'accès de la télomérase au télomère.

III-3-2 Subtélomères

III-3-2-1 Organisation

Situés en amont des télomères, les subtélomères sont des régions variables selon les organismes mais présentant toujours des blocs de séquences répétées (Mefford et Trask, 2002). Typiquement, ils contiennent des répétitions dégénérées du télomère, d'autres types de répétitions et des pseudogènes. Des gènes intacts peuvent être cependant présents et l'importance de ces gènes a été particulièrement révélée chez un nombre conséquent de parasites.

Chez l'Homme, la plupart des minisatellites sont subtélomériques. Des études de génomique comparative indiquent que tous les minisatellites sont probablement d'origine subtélomérique et que la présence de minisatellites en d'autres loci résulte de mécanismes de conversion (Martin et al, 2002).

Chez *S. cerevisiae*, la structure des régions subtélomériques a été bien caractérisée (Louis, 1995). Un « cœur » d'environ 475 pb, appelé élément **X**, adjacent aux répétitions télomériques, a été identifié sur toutes les extrémités chromosomiques. En amont de l'élément X, se trouvent une à quatre copies en tandem d'un élément **Y'** (> 5 kb,) contenant un gène putatif pour une hélicase. Entre ces deux éléments, de courtes séquences répétées en nombre variable sont présentes dans certains subtélomères (**Figure 18**).

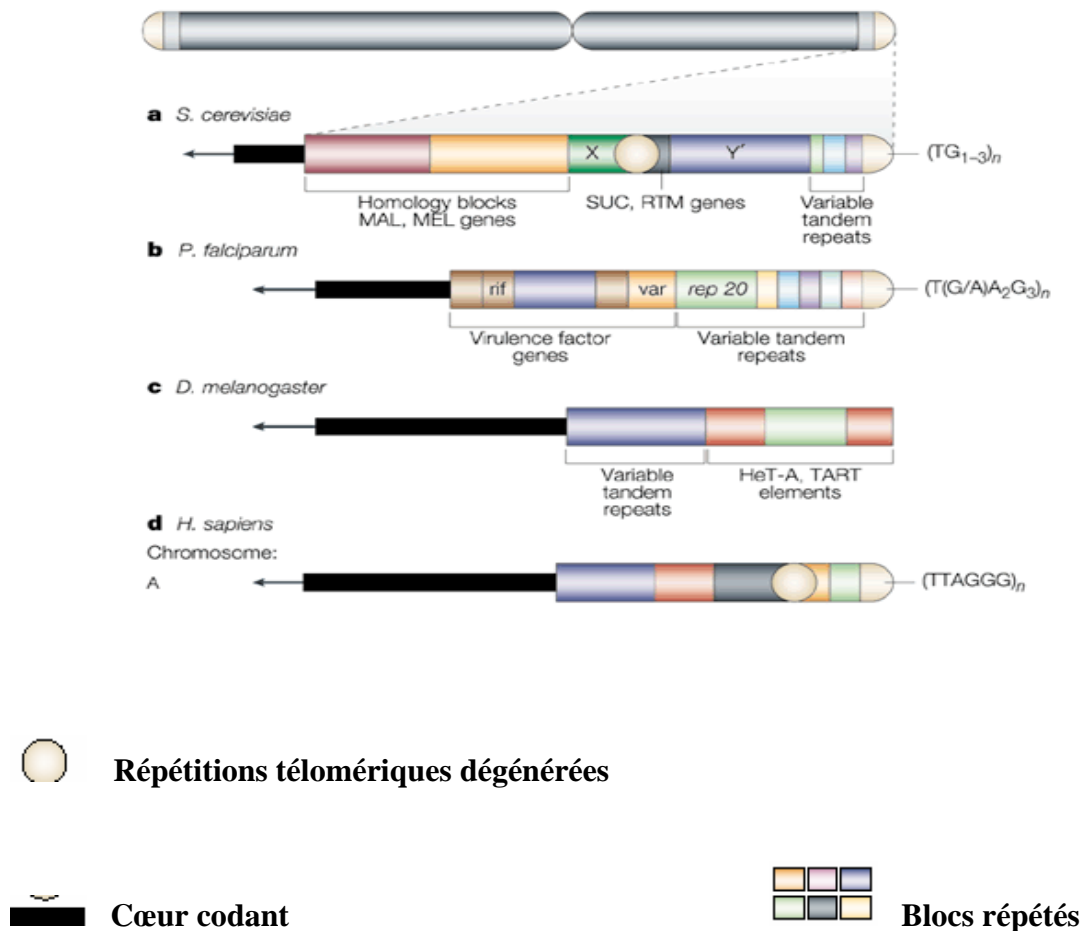


Figure 18 : Structure des subtélomères chez *S. cerevisiae*, *P. falciparum*, *Drosophila melanogaster* et chez *Homo sapiens*. Les subtélomères humains présentent également des familles multigéniques.

La présence de blocs répétés identiques dans les subtélomères de chromosomes non homologues est un fait commun à divers Protistes, tels *Tetrahymena thermophila* (Kirk et Blackburn, 1995), *Giardia lamblia*, (Arkhipova et Morrison, 2001), *Plasmodium falciparum*

(Scherf et al, 2001) et *Trypanosoma brucei* (Bringaud et al, 2002), tout comme à la drosophile (Mason et Biessmann, 1995) et l'Homme (Mefford et Trask, 2002),

III-3-2-2 Importance dans la recombinaison homologue

La recombinaison homologue est définie comme un évènement d'échange (réciproque ou non) entre deux séquences similaires et faisant intervenir l'appareil enzymatique de recombinaison. Ces évènements peuvent avoir lieu au cours de la mitose ou au cours de la méiose, bien qu'ils semblent plus fréquents dans le second cas : il se produit une centaine de recombinaisons au cours de la première division de méiose chez *S. cerevisiae* (Paques et Haber, 1999). On distingue souvent la **recombinaison allélique**, caractérisée par les échanges qui ont lieu entre des chromosomes homologues sur les mêmes loci (c'est la recombinaison "classique"), et la **recombinaison ectopique** qui concerne tous les autres évènements de recombinaison (y compris ceux entre chromosomes hétérologues).

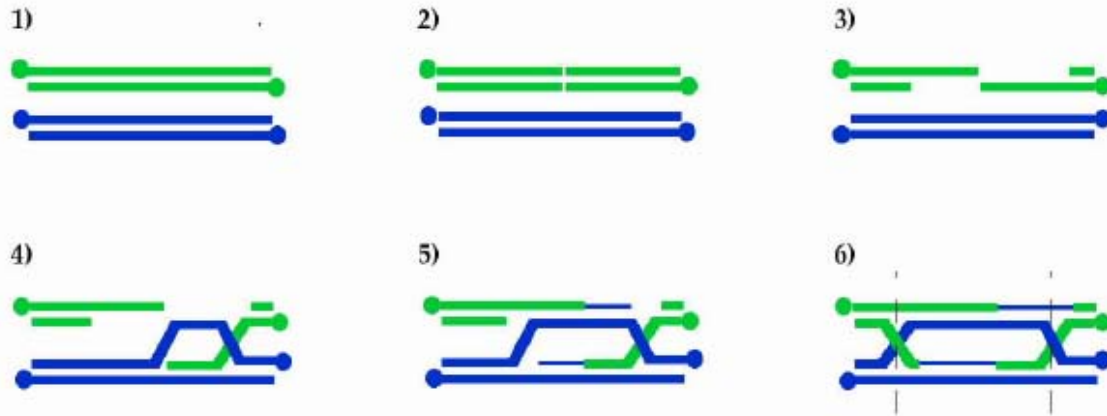
Deux caractéristiques font des subtélomères des endroits favorables à la recombinaison ectopique : (1) le nombre élevé de séquences communes (séquences répétées) aux subtélomères des chromosomes non homologues, (2) le regroupement des télomères à la périphérie nucléaire.

Le modèle aujourd'hui le plus souvent admis pour expliquer la recombinaison homologue, et en particulier la recombinaison durant la méiose, est celui proposé par Szotack et ses collaborateurs (Szostak et al, 1983). Ce modèle rend compte de l'importance de la conversion (échange non réciproque) et du crossing-over (échange réciproque) (**Figure 19**). L'évènement initiateur de la recombinaison est une cassure double-brin. Celle-ci va entraîner la dégradation (de 5' vers 3') des deux extrémités 5' libérées. Puis, les extrémités 3' saillantes vont envahir une séquence homologue et initier une courte synthèse. La ligature de ces néo-brins avec les brins 5' laissés libres va produire une double jonction de Holliday. La résolution de cette structure mènera (ou non) à l'échange réciproque des séquences adjacentes, et la correction des mésappariements entre les deux jonctions de Holliday conduira (ou non) à une conversion.

L'analyse des régions subtélomériques du génome de *S. cerevisiae* a mis en évidence une forte recombinaison ectopique (Britten 1998). Un tel mécanisme appliqué à des séquences codantes pourrait être responsable d'une forte diversification des gènes. Cette hypothèse a conduit à de nombreuses spéculations sur une possible contribution de la recombinaison

ectopique à l'énorme diversité des gènes *OR* (« Olfactory Receptor »), qui forment la plus grande famille de gènes chez l'Homme (Mefford et Trask, 2002).

Les principales étapes :



Les quatre résolutions possibles :

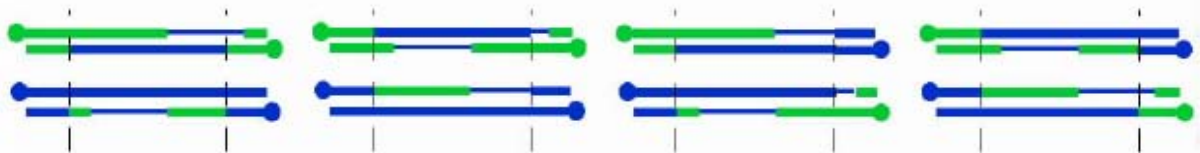


Figure 19: Modèle de recombinaison méiotique.

Deux chromosomes sont représentés avec des couleurs différentes. Les extrémités 5' et les séquences synthétisées sont respectivement symbolisées par des boules et des traits moins épais. Les tirets verticaux délimitent les jonctions de Holliday.

Dans ce modèle, l'évènement déclenchant la recombinaison méiotique est une cassure double-brin (2). Cette cassure entraîne la dégradation 5'→3' des extrémités libres (3), puis l'envahissement de la séquence homologue par les extrémités 3' libres (4). Une courte synthèse (5) mène à deux jonctions de Holliday (6), dont la résolution donne quatre produits possibles (dont deux avec crossing-over)

La comparaison des taux de recombinaison entre séquences subtélomériques et séquences interstitielles chez *S. cerevisiae* a montré que les taux de recombinaisons ectopique et allélique, au cours de la mitose, sont comparables, soit : $2 \cdot 10^{-6}$ événements / cellule / génération (Louis et Haber 1990). Par contre, lors de la méiose, où la recombinaison est fortement influencée par le locus, la recombinaison réciproque entre les régions subtélomériques est réellement diminuée chez *S. cerevisiae*, à l'inverse de ce qui est observé

chez les cellules humaines et chez *Plasmodium* (Freitas et al, 2000). La recombinaison ectopique au niveau des éléments Y' (une famille de répétitions subtélomériques en nombre variable selon les souches de *S. cerevisiae*), semble être cependant limitée aux interactions avec d'autres éléments Y', des phénomènes d'échange ou de conversion avec des régions interstitielles n'ayant pas été détectés (Louis et al, 1994). Il apparaît, donc, que les régions subtélomériques de cette levure ont une propension à recombiner de façon ectopique avec d'autres éléments Y' entre chromosomes non homologues. Contrairement à la diversification fréquemment considérée pour les gènes subtélomériques d'autres organismes, il y a chez *Saccharomyces* une réelle homogénéisation des subtélomères, homogénéisation qui ne se produit pas dans les autres régions du génome.

La recombinaison ectopique des subtélomères permet la réparation des extrémités endommagées mais elle permet également l'échange de séquences. Ainsi, de par leur plasticité, les subtélomères pourraient servir de plateforme à de nouveaux gènes, et donc constituer des lieux de diversification plus rapide. Par exemple, le transfert de séquences entre des gènes paralogues peut créer de nouvelles combinaisons de séquences, variantes éventuellement avantageuses. Des gènes subtélomériques ont effectivement été mis en relation avec des processus d'adaptation chez différents organismes (Barry et al, 2003).

Cette plasticité des régions subtélomériques est à l'origine de divers événements sous-jacents au polymorphisme. Chez *P. falciparum* la présence de répétitions dans le subtélomère pourrait faciliter l'appariement non seulement de chromosomes homologues, mais aussi d'hétérologues, favorisant tant les recombinaisons méiotiques (Hinterberg et al, 1994) que mitotiques (Lanzer et al, 1994a). Les régions subtélomériques sont le siège de cassures double-brin qui constituent une cause majeure de variabilité de taille chromosomique. Elles sont très fréquentes chez *P. falciparum* et concernent vraisemblablement tous ses chromosomes (Lanzer et al, 1994a). La fréquence des cassures dans le gène *Pf11-1*, pendant les mitoses du cycle érythrocytaire du parasite, a été estimée entre 0,03 et 0,15 % (Scherf et al, 1992).

IV- Les petits génomes nucléaires chez les Eucaryotes unicellulaires

Qu'il s'agisse de Protistes ou de Champignons microscopiques, les Eucaryotes unicellulaires présentent souvent un petit génome, la valeur haploïde se situant entre quelques mégabases et une centaine de Mb. Outre les organismes modèles, différents parasites ont été préférentiellement retenus dans les projets de séquençage. Avant de fournir un aperçu de la

diversité génomique, les techniques de cartographie utiles au séquençage systématique et aux analyses comparatives méritent d'être évoquées.

IV-1 Séquençage systématique et nécessité de la cartographie

L'analyse d'un génome peut se faire à diverses échelles de taille, depuis l'observation en microscopie photonique d'un étalement de chromosomes métaphasiques (résolution de plusieurs Mb) jusqu'au stade ultime de détermination de la séquence nucléotidique, soit par dégradation chimique (Maxam et Gilbert, 1977), soit par utilisation de terminateurs de biosynthèse (Sanger et al, 1977). Par la suite, l'automatisation et la robotisation des différentes étapes, depuis la préparation de l'échantillon jusqu'à l'acquisition et l'analyse informatique des données (Cathcart, 1990), ont permis d'accélérer la détermination des séquences d'ADN. Toutefois, la longueur de la séquence dérivée de chaque réaction est limitée à environ 1000 pb. Une stratégie de clonage et d'ordonnancement de différents fragments se chevauchant partiellement est donc requise pour la reconstitution d'une séquence complète. Cet assemblage peut s'effectuer au moyen d'outils informatiques. Les principaux obstacles à l'assemblage sont représentés par la multiplicité des chromosomes et celle des régions répétées dans les génomes complexes, en l'absence d'informations génétiques et/ou physiques sur l'organisation des chromosomes. Il est donc nécessaire de baliser les chromosomes avec des marqueurs ordonnés et construire ainsi des cartes génomiques.

Différents types de cartes sont produits avec des unités de mesure, des marqueurs et des niveaux de résolution différents. Elles sont donc complémentaires et souvent conjointement utilisés pour le repérage des gènes d'intérêt.

- La création de cartes génétiques est basée sur l'analyse des fréquences de recombinaison au cours de la méiose (cartographie méiotique) grâce à des effets phénotypiques déterminables. Les marqueurs phénotypiques sont ordonnés par analyse de leur ségrégation au cours des générations. L'unité de mesure est le centimorgan.

- Des marqueurs moléculaires sont utilisés de façon à augmenter la résolution des cartes génétiques. Ces marqueurs, lorsqu'ils présentent une hétérozygotie, sont analysés comme des marqueurs phénotypiques classiques. Ils peuvent ainsi révéler des polymorphismes:

- de la longueur des fragments de restriction : marqueurs **RFLP** pour « Restriction Fragment Length Polymorphism » (Botstein et al, 1980).

- du nombre de répétitions dans les minisatellites : marqueurs **VNTR** pour « Variable Number of Tandem Repeats » (Nakamura et al, 1987).

- de la longueur des microsatellites : marqueurs **SSLP** pour « Simple Sequence Length Polymorphism » (Su et Wellems, 1996).

-de la longueur de séquences amplifiées au hasard : (**RAPD** pour « Rapid Amplified Polymorphic DNA »).

- La cytogénétique reste une technique d'analyse principalement appliquée aux génomes d'Eucaryotes supérieurs.

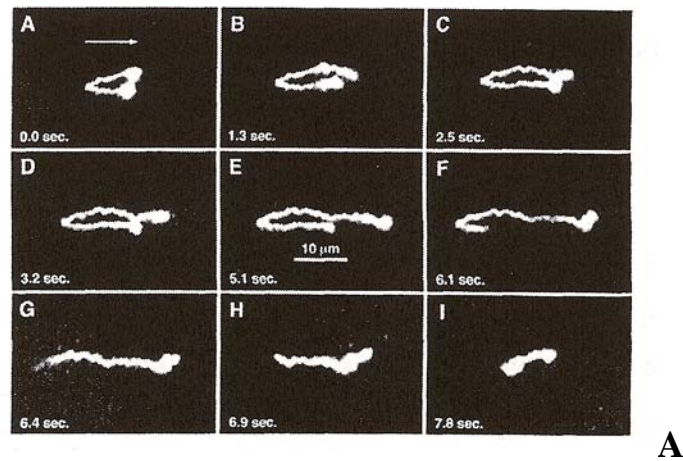
Cartographie physique

Les cartes dites physiques impliquent des marqueurs correspondant à des segments d'ADN chromosomique et l'évaluation des distances réelles entre ces marqueurs. Ces cartes sont d'un très grand intérêt, aussi bien dans le contexte du séquençage systématique que pour des objectifs de génomique comparative.

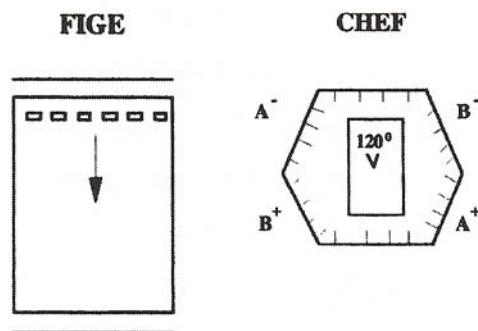
IV-1-1 Cartographie par électrophorèse à champs pulsés

Certaines techniques d'électrophorèse en gel d'agarose, apparues dans les années 1980 permettent la séparation de molécules d'ADN de haut poids moléculaire. Aujourd'hui, les plus utilisées, **FIGE** (« Field Inversion Gel Electrophoresis ») et **CHEF** (« Contour-clamped Homogeneous Electric Field »), dérivent du procédé original de Schwartz et Cantor (1984), abrégé **PFGGE** (« Pulsed Field Gradient Gel Electrophoresis ») ou **PFGE**. Elles reposent sur l'alternance de champs électriques dans des directions différentes, ce qui entraîne des phases de réorientation et de relaxation des molécules d'ADN. Le suivi du comportement d'une molécule d'ADN est montré sur la **Figure 20A**.

La technique FIGE implique l'inversion périodique du champ électrique, les deux électrodes ayant les positions opposées usuelles (Carle et al, 1986). Plus répandue, la technique CHEF assure un champ électrique homogène sur toute la surface du gel, et l'alternance de ce champ s'effectue selon un angle de 120° grâce à un ensemble de 22 à 36 électrodes réparties sur un support hexagonal (Chu et al, 1986) (**Figure 20B**). Cette seconde technique est préférentiellement utilisée pour séparer les grandes molécules d'ADN, jusqu'à 10 Mb (Bustamante et al, 1993). Au-delà, des phénomènes de rétention dans l'agarose sont observés, et leur incidence sur la résolution est d'autant plus néfaste que le voltage et la concentration d'agarose sont élevés (Gurrieri et al, 1999).



A



B

Figure 20: A/ Comportement d'une molécule d'ADN de 164 kb en FIGE. La molécule, colorée avec le composé YOYO-1, a été filmée au cours de sa migration. Elle présente une alternance de phases de compaction et d'élongation. Le moment de la prise de vue est indiqué en bas et à gauche de chaque image (Gurrieri et al, 1996). B/ Schémas illustrant deux types de systèmes PFGE. Le type FIGE permet l'inversion régulière du champ électrique entre deux électrodes parallèles. Dans le type CHEF, l'activation séquentielle de multiples électrodes arrangées hexagonalement permet une alternance de champs électriques formant entre eux un angle de 120°.

Les techniques PFGE, associées au procédé préservant l'intégrité de l'ADN (inclusion de cellules entières dans l'agarose à faible point de fusion avant digestion protéolytique) permettent la séparation de bandes représentatives des molécules d'ADN chromosomiques de différentes tailles. Le profil électrophorétique est alors appelé **caryotype moléculaire**, ce qui sous-entend la correspondance « une bande - un chromosome ». Cette correspondance a été

maintes fois vérifiée chez *S. cerevisiae* dont les 16 chromosomes sont couramment utilisés comme marqueurs de taille en PFGE. La taille du génome haploïde peut alors être calculée à partir de la somme des tailles des ADN chromosomiques. Chez *Plasmodium falciparum*, la microscopie électronique a permis de déterminer 14 chromosomes en comptant le nombre de kinétochores (Prensier et Slomianny, 1986) correspondant bien au nombre de bandes obtenues par PFGE (Kemp et al, 1987). Toutefois, en l'absence de données cytologiques sur le jeu de chromosomes de l'organisme étudié, la prudence s'impose dans l'interprétation du profil PFGE. En effet, deux chromosomes homologues peuvent former deux bandes distinctes s'ils sont de taille différente, de même que deux chromosomes hétérologues peuvent comigrer s'ils sont de taille identique. Un moyen de s'affranchir de ces problèmes est de réaliser des hybridations avec multiples sondes issues du génome considéré, afin de définir les groupes de liaison. Des incertitudes peuvent néanmoins persister pour des organismes au degré de ploïdie inconnu et/ou présentant un grand nombre de chromosomes.

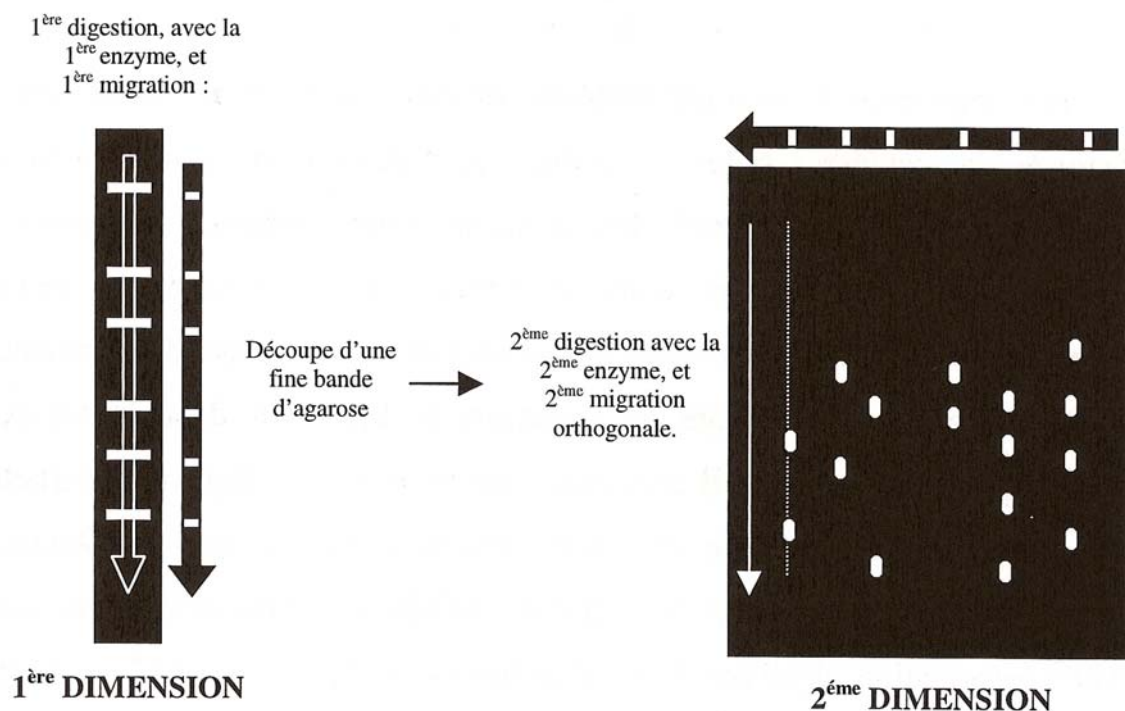


Figure 21: Principe de la cartographie en 2D-PFGE.

Des compléments d'information peuvent être apportés après digestion des chromosomes par une enzyme de restriction à rares sites de coupure (McCelland et al, 1987).

Par exemple, pour les génomes riches en adénine et en thymine, on préférera des enzymes dont le site de reconnaissance à 6 pb est riche en guanine et cytosine (e.g. *Bss*HII, *Bgl*II, *Sma*I etc...). Les fragments de restriction sont séparés dans une 2^e dimension, orthogonale à la première (**Figure 21**). Chaque chromosome présentant un profil de restriction qui lui est propre, la composition complexe de certaines bandes électrophorétiques peut être mieux évaluée en termes de chromosomes homologues et hétérologues (Brugère et al, 2000).

IV-1-2 Fragmentation aléatoire et agencement clonal

Après fragmentation aléatoire du génome, les liens entre les différents fragments sont fréquemment recherchés au moyen de la technique d'agencement clonal, encore appelée ordonnancement de banque (Daniels et Blattner, 1987).

La fragmentation aléatoire est principalement effectuée grâce à des enzymes de restriction à haute fréquence de reconnaissance (par exemple, *Sau*3AI ou *Mbo*I reconnaissant un site de 4 pb), utilisées dans des conditions de digestion partielle. La taille des fragments de restriction est choisie selon la capacité des vecteurs, eux-mêmes choisis selon les besoins de la carte. Les vecteurs les plus simples et les plus anciennement décrits sont des dérivés du phage λ et des cosmides, avec une taille d'insert de 15 à 45 kb. D'autres vecteurs issus du squelette du chromosome artificiel de levure ou YAC (Murray et Szostack, 1983), permettent le clonage de fragments génomiques de 100 à 1000 kb (Burke DT., 1991). Cependant, de nombreux réarrangements, notamment l'apparition de clones chimères (par échange avec le génome « hôte » de *S. cerevisiae*) peuvent être des handicaps dans l'utilisation des YAC, sans compter les difficultés de purification. Ce sont finalement certains vecteurs bactériens, à faible nombre de copies et de capacité plus faible que les YAC, qui sont actuellement préférés pour l'établissement des banques génomiques à ordonner. Ceux-ci sont dérivés soit du facteur F de *Escherichia coli* (vecteurs **BAC**, « Bacterial Artificial Chromosome », d'une capacité de l'ordre de 350 kb, Shiyuza et al, 1992), soit du phage P1 (vecteurs **PAC**, « P1-derived bacterial Artificial Chromosome », d'une capacité de l'ordre de 100 kb, **Sternberg, 1990**).

L'agencement s'opère par recherche d'homologies de séquences entre les différents clones, soit par hybridation, soit par comparaison d'« empreintes » de chaque clone. Afin de réduire le nombre de « trous » dans le génome réorganisé, il est nécessaire d'effectuer ces analyses sur un nombre de clones supérieur à celui déduit par simple division de la taille du génome par la taille moyenne des inserts. Six à huit équivalents de génome seraient

nécessaires pour construire une carte de contigs (Riles et al, 1993). Aujourd'hui, plusieurs étapes sont automatisées et la réorganisation assistée par voie informatique.

IV-2 Exemples de petits génomes

IV-2-1 Chez les Protistes

- Apicomplexes

Ces parasites de divers animaux se présentent sous un état haploïde pendant la majeure partie de leur cycle biologique, le seul stade diploïde étant le zygote issu de la fusion de deux gamètes, chez l'hôte définitif (Cornelissen et al, 1984).

Chez *Toxoplasma gondii*, le génome nucléaire est relativement grand, soit 87 Mb, (Ajioka et al, 2001) réparti sur 11 chromosomes entre 2 et plus de 6 Mb (Sibley et Boothroyd, 1992). Des variations de taille des chromosomes III et V ont été mises en évidence entre différents isolats de l'espèce au cours de cette étude. Le génome de l'agent de la malaria, *Plasmodium falciparum*, est remarquable par sa grande richesse en AT : 82% (Weber, 1987). Sa séquence complète représente 23 Mb pour un total de 14 chromosomes entre 650 kb et 3,5 Mb (Gardner et al, 2002).

D'autres apicomplexes ont un génome plus compact avec un nombre moindre de chromosomes : 9,1 Mb pour 8 chromosomes chez *Cryptosporidium parvum* (Abrahamsen et al, 2004), 8,3 Mb pour 4 chromosomes chez *Theileria parva* (Gardner et al, 2005) et seulement 5,2 Mb pour 3 chromosomes chez *Babesia microti*.

- Kinétoplastidés

Parmi les Flagellés Kinétoplastidés parasites, les leishmanies et les trypanosomes sont particulièrement connus. Les leishmanies présentent des caryotypes très polymorphes, fournissant des évaluations de taille de génome haploïde entre 26 et 67 Mb selon l'espèce considérée (Bastien et al, 1992). Chez *Leishmania major* (souche Friedlin), 36 chromosomes entre 260 kb et 2,8 Mb constituent un génome de ~ 33 Mb (Ivens et al, 2005). Les leishmanies sont généralement considérées comme étant des organismes diploïdes, même si les données actuellement disponibles sont insuffisantes pour l'affirmer de façon catégorique (Bastien et al, 1992).

Dans le cas des trypanosomes, la complexité des caryotypes moléculaires a amené à distinguer trois catégories de taille des chromosomes: les chromosomes « mégabases » (1-6

Mb) qui sont au nombre de 11, les chromosomes « intermédiaires » (200-800 kb) en nombre variable, et les « minichromosomes » (50-150 kb) présents en une centaine de copies par génome (El sayed et al, 2000). Si la diploïdie pour les chromosomes mégabases a été démontrée (Melville et al, 2000), celle pour les chromosomes intermédiaires reste incertaine. Les minichromosomes, constitués à 90 % de séquences de 177 pb répétées en tandem, formeraient un lot aneuploïde (van der Ploeg et al, 1989). Chez le trypanosome africain (*Trypanosoma brucei*), la taille d'un équivalent de génome haploïde serait proche de 35 Mb. Le séquençage complet du lot de chromosomes mégabases a fourni un total de 26 Mb (Berriman et al, 2005). Avec 62 Mb, *Leptomonas collosoma* montre actuellement le plus grand génome parmi les Trypanosomatidés étudiés (Beja et al, 1994). L'analyse de la distribution de certains gènes (gènes d'ARN SL, U6, 5S, 7SL ; ADNr ; gènes de HSP83 et α -tubuline) suggère un noyau diploïde.

- Diplomonadines et Trichomonadines

Dans le groupe des Flagellés Diplomonadida, organismes caractérisés par une symétrie axiale et un état binucléé, *Giardia intestinalis* (ou *lamblia*) est l'espèce la plus étudiée. Son génome haploïde, évalué à 12 Mb, correspond à 5 chromosomes entre 1,5 et 3,8 Mb (Adam, 2000). La forme trophozoïte renferme deux noyaux identiques ; la forme kystique en possède le plus souvent quatre (Adam, 1991). Ces noyaux sont polyploïdes (au moins 4n).

Pour ce qui concerne les Trichomonadida, un premier caryotypage de *Trichomonas vaginalis* avec la technique CHEF avait révélé 6 bandes chromosomiques : 5 bandes entre 1,1 et 5,7 Mb et une bande à seulement 75 kb (Lehker et Alderete, 1999). Toutefois, le nombre réel de chromosomes hétérologues est probablement plus important. Une récente cartographie de ce génome, par restriction des molécules d'ADN dépassant 2 Mb, a fait suggérer une valeur haploïde plutôt surprenante : 180 Mb ! (Upcroft et al, 2006). Le séquençage qui est en cours (<http://www.tigr.org/tdb/e2k1/tvg/>) devrait fournir plus de précisions.

- Entamoeba histolytica

Chez l'amibe dysentérique, l'absence de condensation visible des chromosomes au cours du cycle de développement, la méconnaissance de la ploïdie et la variabilité intra-spécifique de la taille des chromosomes homologues, rendent difficile la détermination du nombre exact de chromosomes. Le polymorphisme de taille chromosomique serait principalement lié à l'expansion-condensation des séquences répétées subtélomériques (Melville et al, 2000). Quatorze groupes de liaison ont été définis à partir de caryotypes

électrophorétiques montrant 31-35 bandes entre 0,3 et 2,2 Mb (Willhoeft et Tannich, 1999). Le séquençage du génome d'*E. histolytica*, a été réalisé au moyen de la technique « shotgun » et indique une valeur haploïde de 24 Mb (Loftus et al, 2005).

IV-2-2. Chez les Champignons unicellulaires

- Hémiascomycètes

Les levures Hémiascomycètes ont un génome très compact, avec une taille généralement inférieure à 20 Mb. Les souches de certaines espèces, notamment *Saccharomyces cerevisiae*, présentent la particularité de pouvoir se maintenir de façon stable aussi bien à l'état diploïde qu'haploïde. L'analyse de la séquence du génome de *S. cerevisiae*, souche S288C (12,5 Mb ; 16 chromosomes entre 230 kb et 1,5 Mb) a révélé un niveau de redondance des ORFs s'élevant à 28 % (Goffeau et al, 1996). Pour justifier une telle redondance génique, certains auteurs ont avancé l'hypothèse d'une duplication ancestrale du génome entier (Wolfe et al, 1997).

De la comparaison des séquences génomiques complètes ou partielles de 13 espèces différentes, il ressort clairement que les duplications et les délétions de segments chromosomiques plus ou moins longs ont constitué un élément moteur important de l'évolution des Hémiascomycètes (Dujon et al, 2004). Le génome de l'espèce marine *Debaryomyces hansenii* (12,2 Mb ; 7 chromosomes entre 1,25 et 2,33 Mb) présente le plus haut degré de redondance (49%) avec une très grande fréquence de gènes dupliqués en tandem. La plus divergente des 13 espèces étudiées est *Yarrowia lipolytica* qui détient aussi le plus « grand » génome, soit 20, 5 Mb réparti entre 6 chromosomes.

Le pathogène humain *Candida albicans* étant complètement dépourvu de forme haploïde, le séquençage de son génome a été effectué par shotgun chez une souche diploïde et hétérozygote (Jones et al, 2004). La valeur haploïde atteint 14,85 Mb pour un lot de 8 chromosomes entre 1 et 3 Mb. L'hétérozygotie est marquée par plus de 62 000 polymorphismes, principalement des substitutions uniques. Onze régions hautement polymorphes ont été identifiées, la plus longue s'étendant sur ~ 9 kb au niveau du locus MTL (« Mating Type-Like »).

- Archiascomycètes

Appartenant au groupe des Archiascomycètes, la levure *Schizosaccharomyces pombe* a fait largement pendant à *S. cerevisiae* de par l'abondance des analyses génétiques et s'est révélé précieuse comme modèle d'étude du cycle cellulaire. Trois grands chromosomes (3,5, 4,6 et 5,7 Mb) constituent son génome haploïde de 13,8 Mb, dont le séquençage a été achevé huit ans après celui de *S. cerevisiae* (Wood et al, 2002). Le nombre total de gènes potentiels codant les protéines s'est révélé plus faible que chez *S. cerevisiae*, 43% de ces gènes contenant des introns. Une étude phylogénétique a montré que des duplications géniques ont eu lieu indépendamment dans les mêmes familles conservées de *S. pombe* et *S. cerevisiae* (Hughes et Friedman, 2003). L'évolution parallèle par duplication de gènes apparaît donc importante pour justifier l'origine des adaptations communes aux différentes lignées d'Ascomycètes.

Initialement placé parmi les Protistes, le parasite unicellulaire *Pneumocystis carinii* responsable d'une grave pneumonie chez les personnes immunodéprimées, est en fait un Ascomycète atypique. Il s'agit d'une entité particulièrement hétérogène, subdivisée en multiples « formae speciales » (*P. carinii* f. sp. *carinii*, f.sp. *ratti*, f.sp. *hominis*, etc...). Une grande diversité caryotypique a été mise en évidence (Hong et al, 1990). Certaines études ont fait considérer un noyau haploïde (~7 Mb) (Wyder et al, 1998). Toutefois, l'hypothèse de la diploïdie ne peut pas être actuellement rejetée, au vu des résultats d'analyse par 2D-PFGE du génome de *P. carinii* f.sp. *carinii* révélant la présence d'homologues pour au moins deux chromosomes (Cornillot et al, 2002). Cette dernière cartographie montre un lot de 17 chromosomes différents (entre 315 et 690 kb), l'estimation de la valeur haploïde devenant 8,4 Mb. Il n'en demeure pas moins vrai que ce génome est significativement plus petit que celui de toutes les levures précitées.

En fait, c'est dans le groupe des Microsporidia, autre exemple d'organismes nouvellement assignés au règne des Fungi, qu'on trouve les génomes nucléaires les plus réduits (moins de 3 Mb chez des espèces du genre *Encephalitozoon*). Nos travaux se rapportant aux microsporidies, les génomes de ces parasites seront commentés plus en détails dans un chapitre spécifique.

IV-2-3 Chez *Dictyostelium discoideum* (Mycetozoa)

Dictyostelium discoideum est un membre très étudié du groupe des Mycétozoaires ou Myxomycètes dont la position taxinomique a fluctué entre protistes, champignons et animaux, en raison d'un cycle montrant une évolution vers la pluricellularité. Chez les Dictyostélidés,

cette pluricellularité est atteinte par agrégation de cellules initialement isolées sous forme amibe, sans passage par un syncytium plurinucléé.

Le génome de *D. discoideum* est haploïde, très riche en AT (~ 78%) et réparti sur 6 chromosomes dont les tailles varient entre 4 et 7 Mb, soit au total 34 Mb (Eichinger et al, 2005). L'ADNr présente la particularité d'être sous forme d'un élément extrachromosomique (90 kb) à raison de 90 copies par noyau (Firtel et al, 1976). Le séquençage du génome a révélé que les extrémités de chaque chromosome portent des copies partielles de cet élément d'ADNr (Eichinger et al, 2005). Les séquences des protéines codées sont riches en longs motifs répétés. Les analyses phylogénétiques ont confirmé l'appartenance au super-groupe des Amoebozoa, plus proche des Opisthokonta (Animaux + Champignons) que des Plantes.

Soulignons finalement que, chez les protozoaires parasites, les comparaisons d'isolats d'une même espèce, issus d'aires géographiques ou d'hôtes différents, ont permis de mettre en évidence une importante variabilité de la taille des chromosomes. Ce polymorphisme n'épargne même pas le génome de 2,9 Mb chez la microsporidie *Encephalitozoon cuniculi* (Biderre et al, 1999b). La plasticité du génome nucléaire est si fréquente que certains auteurs l'ont corrélée aux adaptations caractérisant le mode de vie parasitaire (Lanzer et al, 1995). En fait, dans les taxons où co-existent des espèces libres et parasites, comme c'est le cas chez les Ascomycètes, la variabilité caryotypique est la règle générale (Rustchenko et al, 1993). Il est assez remarquable que quelques mois de culture dans un milieu usuel ont été suffisants pour que 4 des 7 isollements d'une même souche de *S. cerevisiae* présentent un ou plusieurs changements de taille chromosomique (Miklos et al, 1997).

V- Familles multigéniques impliquées dans la variation antigénique chez les parasites

Pour se protéger de l'élimination par le système immunitaire de son hôte, un micro-organisme pathogène peut avoir recours à l'expression séquentielle de différentes versions d'une protéine de surface antigénique, ces variants étant codés par une famille multigénique. Ce phénomène dit de **variation antigénique** a été bien analysé chez certaines bactéries, par exemple chez le spirochète *Borrelia hermsii* où l'expression d'une lipoprotéine de surface dépend de multiples gènes localisés vers les extrémités des plasmides linéaires (Dai et al, 2006). Chez les Eucaryotes, les régions subtélomériques des chromosomes linéaires sont riches en familles multigéniques dont certaines ont été clairement mises en relation avec la

variation antigénique. Les antigènes de surface correspondants sont de nature glycoprotéique et jouent parfois un rôle majeur dans les processus d'adhérence. Nous décrivons ci-dessous quelques familles multigéniques impliquées dans la pathogénie des parasites eucaryotes. Les différents modes de régulation de leur expression seront ensuite examinés.

V-1 Exemples de familles multigéniques

V.1.1. Famille *var* chez *Plasmodium falciparum*

Le cycle des parasites apicomplexes du genre *Plasmodium*, responsables de la malaria (ou paludisme), présente une phase d'invasion des érythrocytes (**Figure 22**), au cours de laquelle la famille de gènes *var* joue un rôle déterminant dans le maintien de la parasitémie. La protéine codée par un gène *var* (entre 200 et 350 kDa, Howard et al, 1983) est dite **PfEMP1** (« *Plasmodium falciparum* Erythrocyte Membrane Protein 1 ») car sa sécrétion est caractérisée par la traversée de la membrane de la vacuole parasitophore puis l'exportation vers la membrane du globule rouge infecté. Elle est ensuite impliquée dans un phénomène de cytoadhérence au niveau des capillaires sanguins, qui mène à la séquestration des hématies et l'obstruction des vaisseaux (Kyes et al, 2001). Les cellules parasitées évitent ainsi le passage par la rate où elles seraient identifiées puis détruites (Miller et al, 1994). L'importance de la séquestration dans certains organes justifie des formes très graves du paludisme, telles que la forme cérébrale (neuropaludisme) et la forme placentaire (paludisme gestationnel).

Un seul variant PfEMP1 est exprimé dans une cellule individuelle. Le changement d'antigène peut s'accompagner d'un soudain changement de virulence (Staalsoe et al, 2002). Le taux de commutation par génération a été estimé à 2% mais, au cours de la phase initiale des infections aiguës, ce taux pourrait atteindre 18% (Gatton et al, 2003). Après insertion dans la membrane érythrocytaire, PfEMP1 peut interagir avec différentes glycoprotéines de l'endothélium vasculaire (Berendt et al, 1990). La plupart des variants sont capables de se lier à la partie « tête » de la molécule du récepteur CD36.

L'annotation du génome de *P. falciparum* a permis de mieux préciser le répertoire de gènes *var*. Ainsi, chez la souche de référence 3D7, 59 gènes *var* intacts co-existent avec des fragments de gènes et des pseudogènes (Gardner et al, 2002). Des groupes de gènes (« clusters ») répétés en tandem sont localisés dans les régions internes de plusieurs

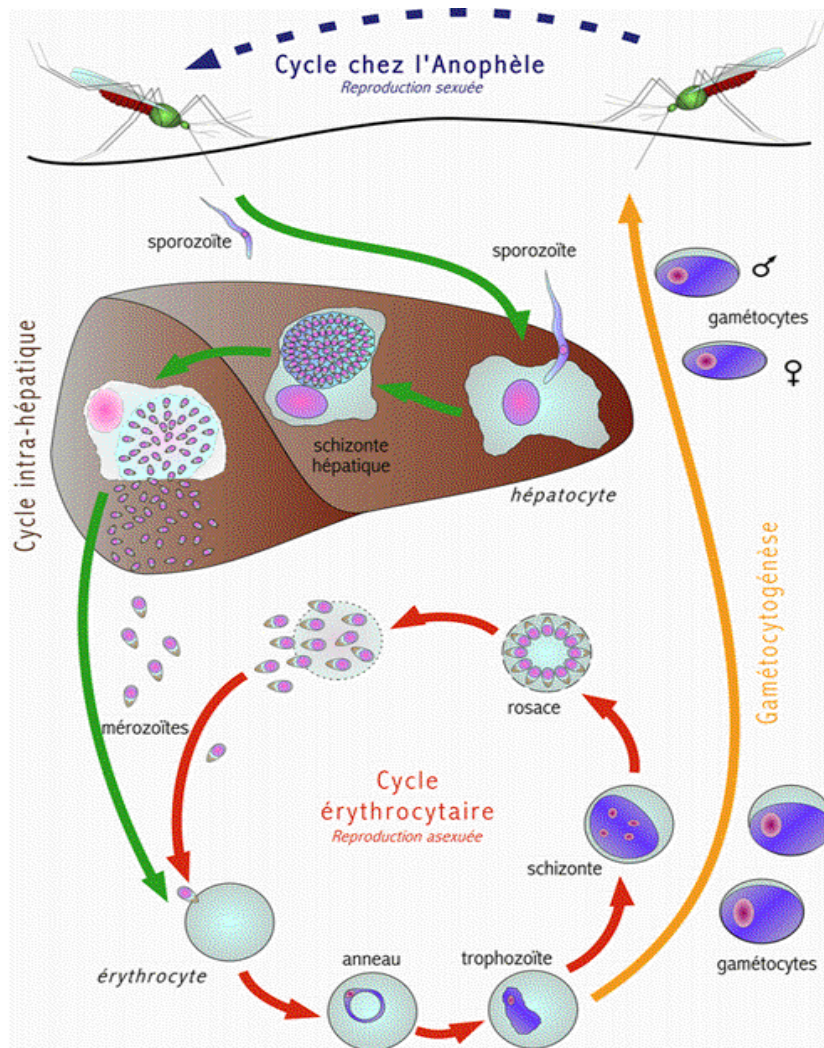


Figure 22: Cycle de développement de *Plasmodium* sp.

chromosomes. D'autres gènes sont associés aux régions subtélomériques de la plupart des chromosomes, à raison de 1-3 copies par subtélomère (**Figure 23B**).

Tous les gènes *var* ont la même structure de base, à savoir deux exons séparés par un intron conservé (**Figure 23A**) :

- le premier exon code la partie extracellulaire de la protéine et un domaine transmembranaire simple. La séquence de la partie extracellulaire est hypervariable mais présente des domaines faiblement similaires au domaine de liaison « Duffy » de l'antigène de surface identifié chez *P. knowlesi* et *P. vivax* (Adams et al, 1992), d'où l'appellation domaines DBL (« Duffy Binding-Like ») (Su et al, 1995). La variabilité de l'exon 1 est à l'origine des différences d'affinité de liaison et d'antigénicité entre les variants.

- le second exon code une région fortement chargée qui servirait d'ancrage à la protéine au niveau des petites protubérances (« knobs ») de la surface érythrocytaire, impliquées dans la séquestration (**Figure 24**). Contrairement à l'exon 1, l'exon 2 est fortement conservé au sein de la famille (Su et al, 1995).

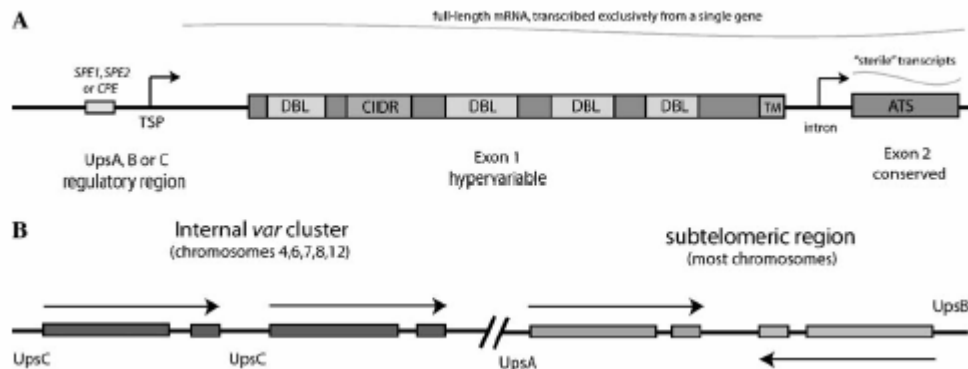


Figure 23: Structure et arrangement des gènes *var*. **A/** Structure classique d'un gène *var*. Les flèches marquent les promoteurs. L'exon 1 code pour plusieurs domaines DBL. L'exon 2, région la plus conservée, code pour un segment terminal acide. **B/** Disposition des gènes *var* à l'intérieur du génome. Des clusters de gènes *var* avec un promoteur de type UpsC ont souvent une localisation interne tandis que ceux avec des promoteurs de type UpsA ou UpsB sont typiquement subtelomériques (**Frank et al, 2006**).

La plupart des protéines codées possèdent également une région riche en cystéine (CIDR pour « Cystein-rich Inter-Domain Region ») qui, avec les domaines DBL, contribuent à la cytoadhérence des cellules infectées aux molécules de surface de l'endothélium vasculaire (Baruch, 1999). La sécrétion de Pfemp1 et de certaines autres protéines à travers la membrane de la vacuole parasitophore a pu être corrélée à une séquence-signal particulière, ce qui indique une reconnaissance par un système récepteur-transporteur (Marti et al, 2004).

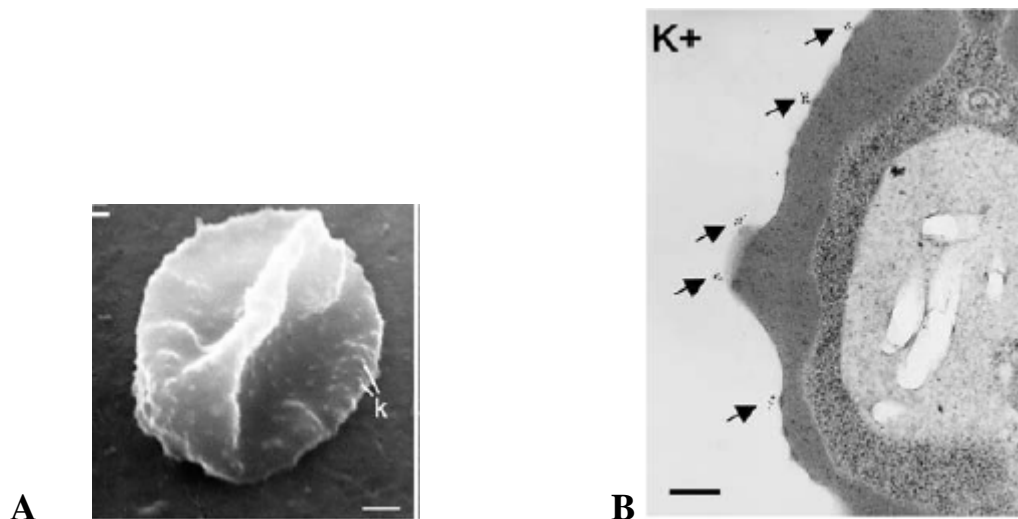


Figure 24 : *A/ « Knobs » (k) à la surface d'une hématie parasitée par Plasmodium falciparum, observée en microscopie électronique à balayage. B/ Localisation immunocytochimique à l'or colloïdal de la protéine PfEMP1. La surface de l'hématie parasitée est marquée au niveau des « knobs » (flèches)(Horrocks et al, 2005)*

V.1.2. Famille *vsg* chez *Trypanosoma brucei*

Les études sur les protéines de surface des trypanosomes africains ont été dominées par les deux familles suivantes:

- les **VSG** (« Variant Surface Glycoprotein ») chez la forme « bloodstream » ou « sanguine » qui vit dans le système circulatoire de l'hôte,
- les **procyclines** ou **PARP** (« Procyclic Acidic Repetitive Protein ») chez la forme « insecte » ou « procyclique » qui se développe dans le tube digestif de l'insecte vecteur, la mouche tsé-tsé ou glossine (**Figure 25**).

Ces deux types de glycoprotéines, spécifiques de stades de développement, couvrent la surface entière du trypanosome. Le manteau de VSG est acquis quand les parasites ont terminé leur maturation dans les glandes salivaires du vecteur, c'est-à-dire pendant la phase métacyclique. Il persiste durant toute la période de développement dans la circulation sanguine de l'hôte-mammifère, puis il est remplacé par un manteau de procyclines quand les trypanosomes ingérés par la glossine se différencient en formes procycliques.

La séquence presque complète du génome de *T. brucei* révèle la présence d'environ 1700 copies de gènes *vsg* dont la plupart sont des pseudogènes (Berriman et al, 2005)

(<http://www.genedb.org/genedb/trypan>). La grande majorité de ces copies sont groupées en tandem dans les subtélomères des chromosomes. D'autres copies sont très proches des

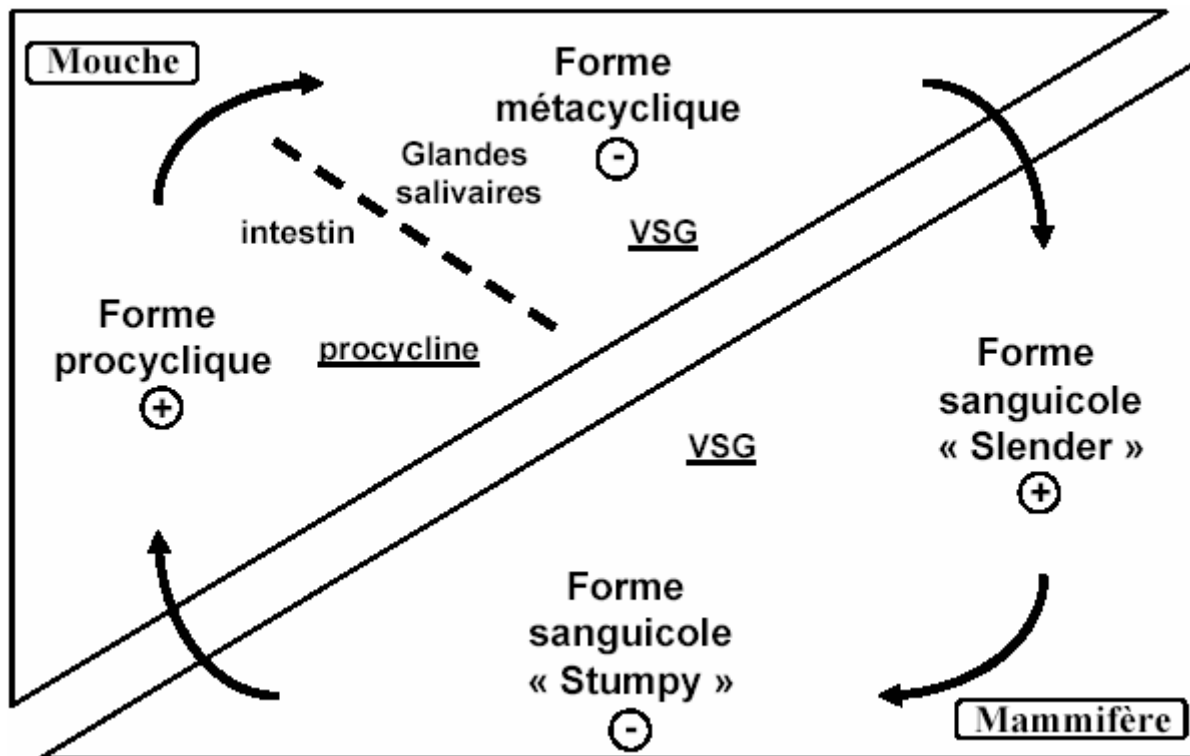


Figure 25: Représentation schématique des étapes importantes du cycle de vie de *Trypanosoma brucei*. Le cycle est une alternance d'étapes de prolifération (+) et de quiescence (-). L'antigène de surface (VSG ou procycline) est indiqué pour quatre formes parasites principales. Ce schéma très simplifié ne rend pas compte des formes intermédiaires.

télomères (position dite télomérique), en particulier sur les minichromosomes d'environ 50-150 kb, mais également à l'extrémité des chromosomes mégabases (Rudenko et al, 1998). Pour certains auteurs, la fonction des minichromosomes, au nombre de 100 environ, serait de fournir au parasite un important réservoir de *vsg*. Pour être exprimé, le gène doit se trouver dans un locus particulier, appelé site d'expression. Ce locus, d'une taille de 50 kb, est polycistronique et occupe une position télomérique sur les chromosomes mégabases (Pays et al, 2001) (**Figure 26**). Au total, il existe environ 15 sites fonctionnels par génome (Becker et al, 2004).

La protéine VSG, d'environ 60 kDa, se compose d'un domaine N-terminal de 350 à 400 aa et d'un domaine C-terminal de 50 à 100 aa. Elle est accrochée à la membrane

plasmique par une ancre GPI (glycosyl-phosphatidylinositol), près de l'extrémité C-terminale et contenant deux résidus d'acide myristique. Les molécules s'associent en dimères qui adoptent une configuration perpendiculaire à la surface cellulaire, due au repliement du domaine N-terminal en deux longues hélices α antiparallèles séparées par un coude (Blum et al, 1993). La comparaison des séquences VSG provenant de différents isolats révèle une importante variabilité du domaine N-terminal. Par contre, le domaine C-terminal est relativement conservé, particulièrement autour des résidus cystéine qui forment des ponts disulfure. Toutes les VSG décrites sont N-glycosylées au niveau de ce dernier domaine. Les variations de séquence n'ont pas d'incidences notables sur la structure tertiaire.

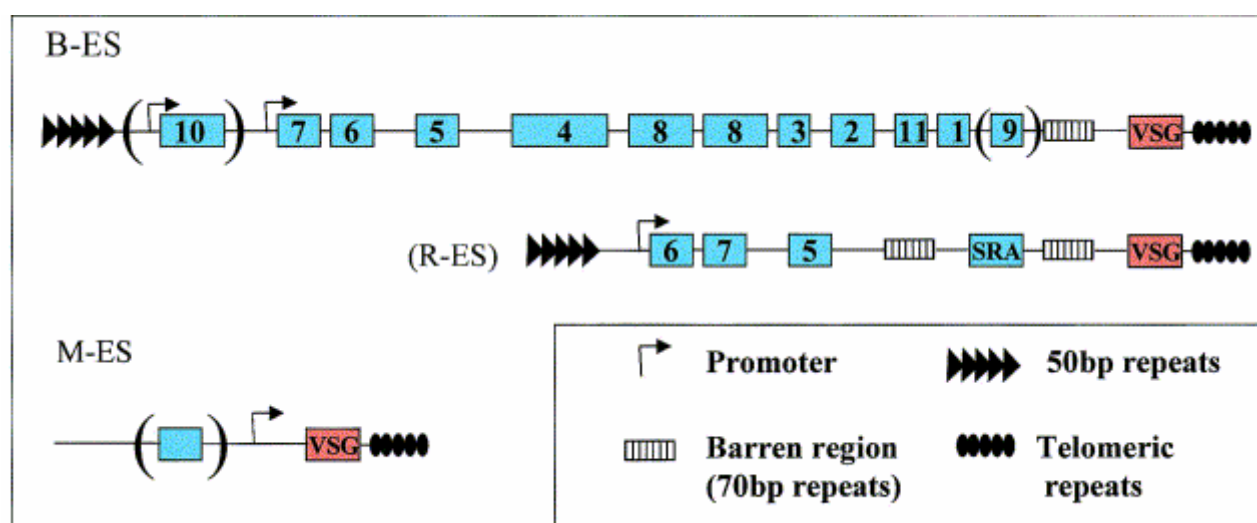


Figure 26: Représentation schématique de sites d'expression de *vsg* (Pays et al, 2001). Chez les formes sanguines, les sites usuels (B-ES) sont des unités polycistroniques contenant un seul *vsg* (en rouge) et plusieurs ESAG (« Expression Site Associated Genes ») (en bleu et numérotés). Toujours situé en fin d'ES, le *vsg* est bordé en aval des répétitions télomériques et en amont de répétitions de 70 bp. L'ES est bordé en amont par des répétitions de 50 bp. Les éléments entre parenthèses ne sont pas toujours présents. Le variant R-ES est moins long et caractérisé par un gène SRA (« serum-resistance-associated »), impliqué dans la résistance à la lyse par le sérum humain. Chez la forme métacyclique, les sites d'expression (M-ES) sont réduits à des unités monocistroniques, quelques copies dégénérées d'ESAG pouvant se trouver en amont.

La fonction principale des VSG est de former un manteau protecteur très dense. Chaque cellule est entourée par 10^7 molécules d'un variant VSG donné (Ziegelbauer K et Overath P, 1993). Ces molécules sont si étroitement liées que seule une partie très limitée de

leur séquence d'acides aminés est accessible à l'environnement. Cette organisation empêche non seulement la reconnaissance des épitopes conservés de la région C-terminale des VSG, mais elle protège également les protéines invariables de la surface cellulaire qui sont potentiellement vulnérables aux réactions du système immunitaire. Toutefois, les épitopes superficiels des VSG restent aisément détectables par l'hôte, et la réponse humorale résultante mène à l'élimination efficace et rapide de nombreux parasites dans la circulation sanguine. L'expression d'un nouveau variant VSG permet à quelques membres de la population parasitaire d'échapper aux anticorps dirigés contre le variant précédent. L'effet continu entre la réaction immunitaire de l'hôte et la variation antigénique du parasite donne un schéma typique de parasitémie avec une succession de pics représentatifs de populations de trypanosomes qui diffèrent seulement par leur VSG de surface (**Figure 27**). L'avantage de cette stratégie est qu'elle assure une infection persistante avec un nombre tolérable de parasites dans le sang, sans entraîner une mort rapide de l'hôte (Pays et Nolan 1998).

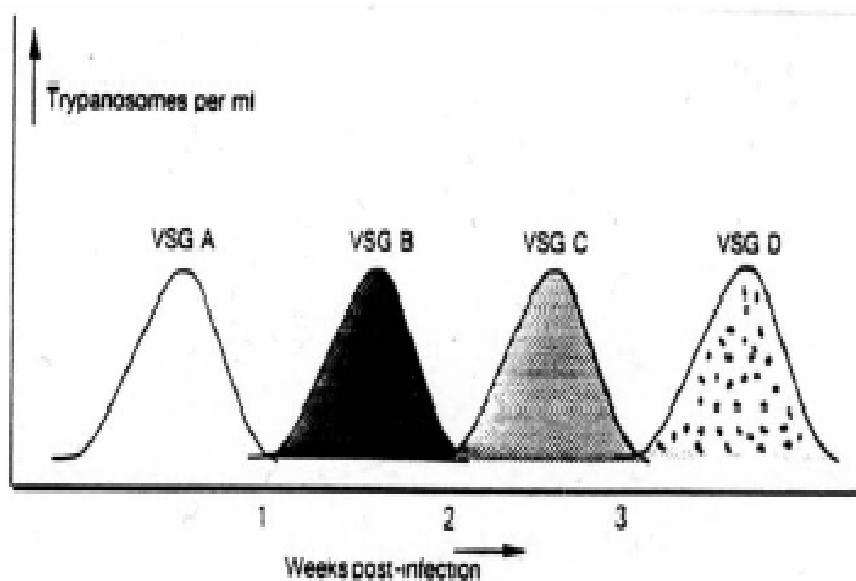


Figure 27: Représentation graphique des vagues de parasitémie développées par une succession de populations de trypanosomes dans la circulation sanguine, exprimant chacune un *vsg* différent (1, 2, 3 etc...). Chaque pic correspond à une population de parasites antigéniquement homogène.

V.1.3. Famille *vsp* chez *Giardia intestinalis* (*lamblia*)

La diplomonadine *Giardia intestinalis* est l'un des principaux protozoaires responsables d'infections intestinales chez l'Homme. Une grande diversité moléculaire chez

les isolats de cette espèce réside au niveau de protéines de surface, dites **VSP** (« Variant-specific Surface Protein »), qui forment un manteau couvrant la surface entière du trophozoïte (Nash, 1992). Bien que les VSP puissent varier considérablement en séquence et en taille (entre 50 et 250 kDa), elles ont en commun une teneur élevée en cystéine (~ 12%) et en thréonine (~ 11%), multiples copies d'un motif tetrapeptidique (CXXC) et une séquence C-terminale fortement conservée (Ey et al, 1993). La survie du trophozoïte dans l'intestin du vertébré semble bien dépendre d'une variation antigénique impliquant les VSP (Nash, 1997).

Le répertoire de gènes *vsp* a été estimé entre 133 et 151, ce qui représente ~ 2% du génome haploïde (Nash et Mowatt, 1992). Cependant, contrairement aux familles précédemment décrites, ces gènes ne sont pas particulièrement localisés aux extrémités chromosomiques (Yang et Adam, 1994). La variation antigénique se présente comme un processus continu, les transcriptions successives d'au moins 22 gènes *vsp* différents ayant été démontrées au cours d'une infection de 42 jours avec le clone GS/M-83-H7 (Bienz et al, 2001). L'enkystement fait intervenir l'expression d'un nouveau variant, mais simultanément avec d'autres VSP (Carranza et al, 2002).

V.1.4. Exemples de familles chez des champignons

Les champignons présentent diverses familles de gènes codant des glycoprotéines de surface qui confèrent à la paroi cellulaire des propriétés d'adhésion et d'immunogénicité particulières. L'adhésion peut concerner une surface abiotique ou la surface d'autres cellules (revue Verstrepen et Klis, 2006). Les protéines impliquées dans cette fonction sont nommées soit adhésines, soit floculines. Chez *S. cerevisiae*, les protéines codées par des familles **FLO** confèrent une adhérence à l'agar et aux surfaces solides mais permettent également l'agrégation calcium-dépendante des cellules de cette levure, justifiant le phénomène de floculation (Guo et al, 2000). Des familles multigéniques pour des protéines responsables de l'adhérence aux tissus de mammifères ont été caractérisées chez des espèces parasites telles que *Candida albicans* (gènes **ALS et EAP**) et *C. glabrata* (gènes **EPA**) (Cormack et al, 1999 ; Hoyer 2001). Plusieurs de ces familles sont localisées près des télomères. Deux types d'adhésion ont été décrits chez les levures :

- une adhésion sucre-dépendante, faisant intervenir des adhésines de type lectine dont la région N-terminale possède un domaine de liaison au sucre. Les gènes **FLO** (excepté **FLO11**) de *S. cerevisiae* et les gènes **EPA** de *C. glabrata* déterminent de telles lectines.

- une adhésion sucre-indépendante, impliquant des adhésines qui reconnaissent des peptides à la place des sucres. C'est le cas de FLO11 de *S. cerevisiae* ainsi que des protéines ALS de *C. albicans* qui interagissent avec certains peptides de la cellule-hôte.

La famille *msg* codant pour une glycoprotéine majeure de surface chez *Pneumocystis carinii* (Kovacs et al, 1988) a aussi retenu l'attention. Ayant une taille variant entre 90 à 140 kDa, les multiples formes de MSG sont hautement immunogènes. Celles-ci sont impliquées dans les réponses de l'immunité aussi bien cellulaire et qu'humorale de l'hôte (Linke et al, 1998). La glycoprotéine MSG serait un récepteur de la fibronectine et permettrait ainsi l'attachement aux pneumocytes (cellules alvéolaires) (Pottratz et Martin, 1990). Au niveau histologique, l'infection se caractérise par un recouvrement des parois alvéolaires par de nombreux trophozoïtes adhérant aux cellules épithéliales.

Le nombre de gènes *msg* n'est pas encore clairement défini. Des expériences d'hybridation ont conduit à une évaluation de 100 gènes par génome haploïde chez une souche de *P. carinii* (Sunkin et al, 1994). Ces gènes sont localisés au niveau des régions subtélomériques de tous les chromosomes, formant des blocs de 2-4 gènes (Sunkin et al, 1996).

V-2 Régulation de l'expression des familles multigéniques

Bien que le but soit d'exprimer un seul gène à un moment donné (« expression mutuellement exclusive »), les mécanismes qui régissent l'expression des familles multigéniques sont multiples (Borst et al 1998).

V-2-1 Promoteurs et sites d'expression

L'initiation de la transcription constitue le premier niveau de contrôle de l'expression de ces gènes. L'activité des promoteurs est donc à considérer mais la localisation du gène sur le chromosome peut éventuellement être aussi déterminante. Ceci nous amène à commenter trois cas distincts.

V-2-1-1 Gènes *var* : trois types de promoteurs

Chez *Plasmodium falciparum*, le site d'initiation de la transcription d'un gène *var* donné est situé approximativement à 1 kb en amont de l'ORF, ce qui indique que la région

non traduite 5' est plutôt longue (Deitsch et al, 1999). Trois types de séquences promotrices (**UpsA**, **B** et **C**) ont été caractérisés, reflétant une localisation différentielle des gènes *var* (Lavstsen et al, 2003). Les gènes associés à UpsA sont exclusivement trouvés dans les régions subtelomériques, orientés vers le télomère, tandis que ceux associés à UpsC sont localisés dans les clusters internes. Les gènes avec UpsB sont répartis dans les subtelomères et les clusters internes. La signification de ces différents types de promoteurs n'est pas entièrement comprise. Il a été envisagé qu'ils serviraient au contrôle de l'expression de différentes sous-classes de gènes *var*, ou qu'ils seraient le résultat d'une recombinaison sélective entre les membres de ces sous-classes pendant la génération de la diversité au sein de la famille (revue Frank et Deitsch, 2006). D'autres séquences régulatrices potentielles ont été révélées par la technique EMSA (« electromobility shift assay ») appliquée à des extraits nucléaires (Voss et al, 2003). Situées en amont des promoteurs UpsB et UpsC, elles se lient aux complexes protéiques issus de noyaux de stades tardifs du parasite qui sont des stades où les gènes *var* ne sont pas transcrits. Ce résultat soulève donc l'hypothèse d'un rôle dans des modifications épigénétiques conditionnant l'activation ou l'inactivation de gènes au cours des cycles cellulaires successifs.

L'activité des promoteurs *var* a été testée par transfection en utilisant des constructions dans lesquelles les régions régulatrices de divers gènes *var* ont été clonées en amont de gènes rapporteurs codant la luciférase ou la chloramphénicol acétyltransférase (Deitsch et al, 1999 ; Voss et al, 2000). Une transcription des gènes rapporteurs a été observée pour toutes les séquences promotrices testées, et des expériences de délétion ont permis d'identifier les séquences minimales requises pour l'activité transcriptionnelle. Les résultats obtenus suggèrent aussi que le déplacement de ces éléments régulateurs en dehors de leur contexte chromosomique perturbe leur capacité d'inhibition.

Des transfections impliquant le placement d'un intron *var* en aval du gène rapporteur, ont montré un très fort effet inhibiteur de cet intron sur l'activité d'un gène témoin placé sur le même plasmide qu'un promoteur *var* (Deitsch et al, 2001). L'inhibition ne dépend pas de l'orientation de l'intron mais exige le passage des parasites transformés par au moins une phase S (synthèse d'ADN au début du stade trophozoïte), ce qui constitue une caractéristique du « silencing » des gènes par modification chromatinienne (Li et al, 2001). Des travaux supplémentaires sur les séquences introniques *var* et leurs propriétés inhibitrices ont permis de définir trois régions distinctes sur la base du contenu en bases GC et de l'asymétrie de la

molécule d'ADN (Calderwood et al, 2003). La région centrale assure la fonction promotrice de l'intron, avec une efficacité plus forte lorsque les transcrits « stériles » deviennent plus abondants au cours du cycle (Kyes et al, 2003). Il est donc probable que l'intron soit la source de ces ARN non codants. La création d'une construction portant un intron sans sa région promotrice a pour conséquence la perte de son pouvoir inhibiteur, démontrant ainsi la nécessité de ce promoteur (Calderwood et al, 2003). Ainsi, des interactions entre les deux promoteurs trouvés dans chacun des gènes *var* (l'un en amont de la région codante et l'autre dans l'intron), semblent contribuer à l'inactivation de ces gènes.

Une régulation par des ARNm anti-sens est envisageable mais une inhibition consécutive à l'augmentation du niveau des transcrits anti-sens n'a pu être démontrée expérimentalement (Ralph et al. 2005).

V-2-1-2 Gènes *vsg* : plusieurs sites d'expression avec un promoteur de type ribosomique

L'expression des gènes *vsg* chez *Trypanosoma brucei* semble plus complexe dans la mesure où le gène actif doit se trouver dans un locus génomique particulier, appelé **site d'expression (ES)**. Plusieurs ESs sont présents dans le génome. Leur position est télomérique sur les chromosomes mégabases et le nombre total de télomères portant un ES a été estimé à 15 par génome haploïde (Pays et al, 2001, Becker et al, 2004). D'une taille de 50 kb, chaque ES est représenté par une unité de transcription polycistronique où le gène *vsg* est précédé par une collection variable de gènes **ESAG** (« Expression Site Associated-Genes », 10 gènes au maximum) (**Figure 26**). Deux de ces gènes (ESAG6 et ESAG7) sont proches du promoteur et communs aux différents ES, les autres étant en nombre et position variables. Les protéines de surface codées par les ESAG sont pour la plupart indéfinies. ESAG4 coderait un récepteur transmembranaire de type adénylate cyclase, alors qu'ESAG6 et ESAG7 coderaient les deux sous-unités d'un récepteur hétérodimérique à la transferrine de l'hôte, nécessaire à la croissance du parasite. Conférant à *T. brucei rhodesiense* sa capacité de résister à l'attaque du sérum humain, le gène SRA (« Serum Resistance-Associated ») a été identifié dans un seul ES, dit R-ES (Xong et al, 1998). De nombreux gènes liés aux ESAGs (GRESAG pour « Genes Related ESAG ») sont transcrits indépendamment.

La comparaison de plusieurs promoteurs d'ES a montré un haut pourcentage d'identité de séquence (Pays et al, 1990). L'expression d'un gène rapporteur sous contrôle d'un promoteur d'ES a été observée, que l'ES soit actif ou inactif dans le génome, même au stade

procyclique où normalement le gène *vsg* n'est pas exprimé. L'analyse par délétion de séquences a montré que les 70 pb en amont du site d'initiation de la transcription sont suffisantes pour une activité optimale du promoteur (Jefferies et al, 1991). Deux motifs essentiels sont appelés boîtes 1 et 2. Les résultats d'insertions et délétions de nucléotides entre ces deux boîtes suggèrent une distance optimale pour le bon fonctionnement du promoteur. Des séquences équivalentes aux boîtes 1 et 2 sont présentes dans les régions promotrices des gènes ribosomiques (Zomerdijk et al, 1991). Le promoteur pour les gènes de procyclines a une organisation similaire (**Figure 28**). Un promoteur chimérique dans lequel la deuxième moitié du promoteur *vsg* est remplacée par son équivalent dans le promoteur ribosomique est tout à fait opérationnel (Vanhamme et al, 1995b). Le promoteur d'ES peut même être remplacé *in situ* par un promoteur ribosomique sans altérer la variation antigénique (Rudenko et al, 1995). Si l'on ajoute à ces données l'insensibilité à l' α -amanitine de la transcription d'ES (Kooter et Borst 1984), il peut alors être considéré que le promoteur d'ES est de type ribosomique, c'est-à-dire reconnu par l'ARN polymérase de type I (Pol I). La transcription d'ES par Pol I a effectivement été clairement démontrée dans un système *in vitro* (Günzl et al, 2003).

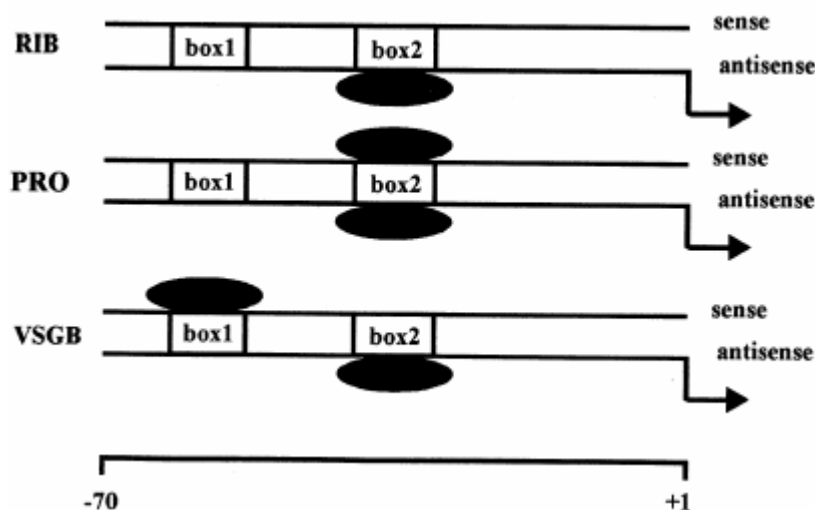


Figure 28: Analyse comparative des régions promotrices des gènes d'ARNr (RIB), de procyclines (PRO) et de VSG (VSGB). Deux boîtes sont essentielles pour la transcription dans les trois cas (pour plus d'explications, voir le texte). (Vanhamme et Pays, 1998)

V-2-1-3 Gènes *msg* : un seul site d'expression incluant une région transcrite commune à tous les ARNm

Le nombre de gènes *msg* exprimés à un moment donné par une cellule de *Pneumocystis carinii* est difficile à évaluer directement en raison des difficultés d'obtention de populations clonales pour ce parasite.

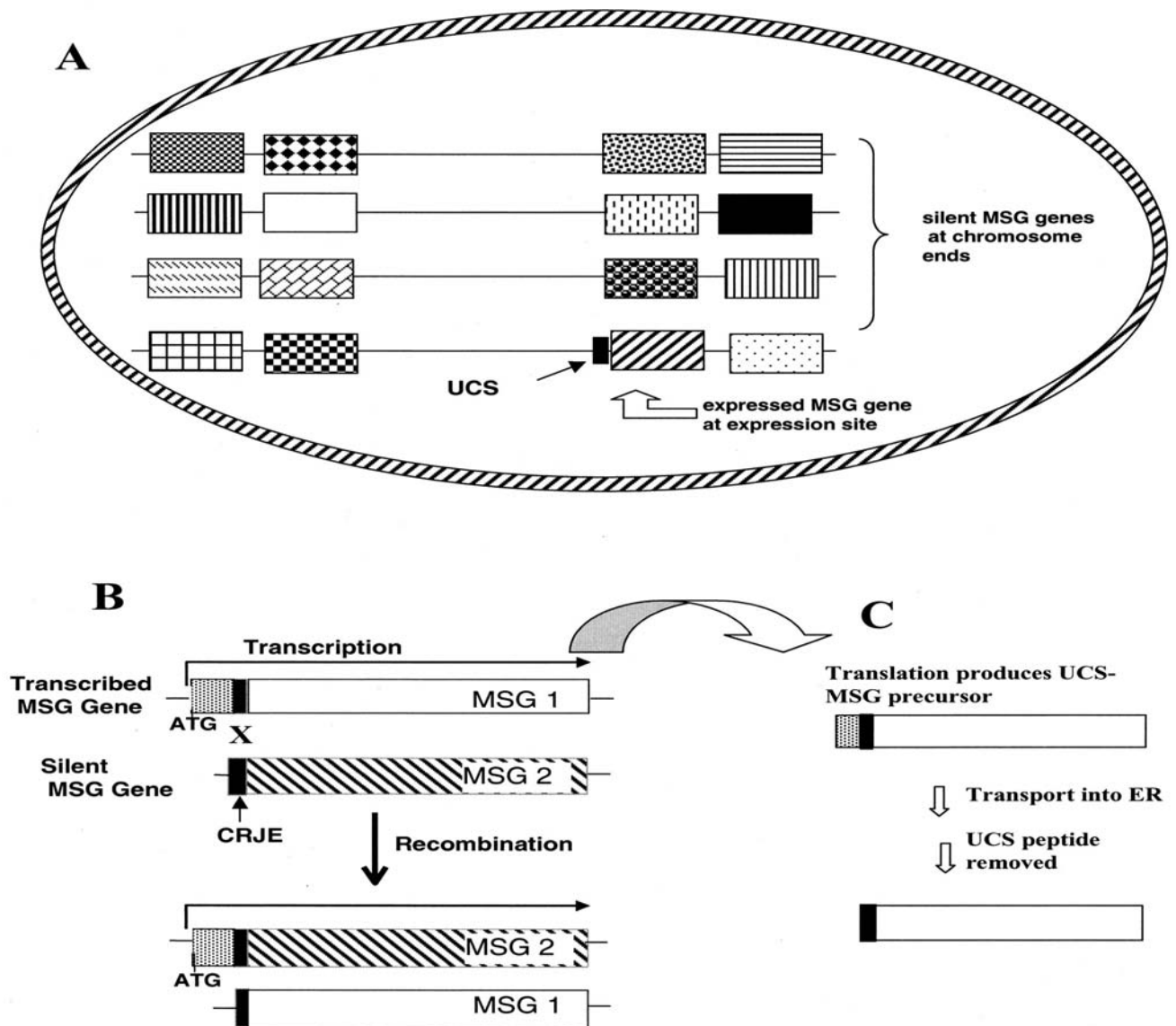


Figure 29: Organisation et expression des gènes *msg* chez *Pneumocystis carinii*.

A/ Les gènes *msg* (rectangles) sont localisés vers les extrémités des chromosomes. Seul le gène au niveau du site UCS est transcrit. B/ Remplacement d'un gène par recombinaison homologue impliquant la séquence CRJE. C/ Après transcription du gène au locus UCS, la séquence UCS-MSG est traduite puis clivée pour donner la protéine finale (Stringer et Kelly, 2001).

La comparaison des séquences d'ARNm codant plusieurs variants MSG a révélé que la région 5' terminale était fortement conservée sur environ 500 pb, d'où l'appellation UCS pour « Upstream Conserved Sequence » (Wada et Nakamura, 1996). Les expériences d'hybridation montrant une représentation de cette UCS au niveau d'une seule bande du caryotype moléculaire, il est apparu probable que la séquence était unique dans le génome. Le nombre de clones portant UCS dans les banques génomiques est effectivement faible et le séquençage de ces clones a montré régulièrement la présence d'un gène *msg* juste en aval de l'UCS. L'absence de polymorphisme dans la carte de restriction de la région amont (~ 10 kb) a finalement fait conclure à un site unique d'expression, en position télomérique.

Dans le modèle proposé (**Figure 29**), la grande majorité des gènes *msg* sont organisés en clusters transcriptionnellement silencieux situés aux extrémités de chacun des chromosomes, et seul le gène *msg* lié au locus UCS est transcrit. Le variant protéique correspondant est alors synthétisé puis exporté à la surface du parasite. Le variant sera changé après remplacement du gène *msg* au locus UCS par un autre. Il est émis l'hypothèse qu'une séquence de 23 pb commune aux *msg* (**CRJE** pour « Conserved Recombination Junction Element ») servirait de cible à des recombinaisons homologues au niveau du site d'expression (Stringer et Kelly, 2001). Outre celle concernant le nombre d'ES, une différence majeure entre *Pneumocystis* et *Trypanosoma* réside dans le fait que l'ES de *Pneumocystis* procure à la fois le promoteur pour la transcription d'un gène *msg* et une séquence transcrite (UCS) incluant le début de la séquence pour la traduction en protéine (et probablement aussi pour la maturation protéolytique), alors qu'un ES de *Trypanosoma* ne sert qu'à la transcription.

V-2-2 L'exclusion allélique

L'exclusion allélique se caractérise par l'expression d'un seul gène au sein d'une famille multigénique à un moment donné, tous les autres étant réprimés. Les mécanismes de contrôle intervenant dans ce phénomène sont de deux types (Borst, 2002) : contrôle par recombinaison de l'ADN (inversion, recombinaison réciproque ou conversion génique) et activation *in situ*.

V-2-2-1 Contrôle par recombinaison de l'ADN

Il s'agit d'un déplacement de gène grâce à un processus de recombinaison de l'ADN activant un promoteur.

La recombinaison peut entraîner une **inversion** du gène (**Figure 30A1**) qui est catalysée par une ADN recombinase reconnaissant les petites séquences flanquant le segment à inverser (van de Putte et Goosen, 1992). Chez les bactéries *Salmonella* faisant varier l'expression de leurs flagellines codées par deux gènes différents H1 et H2, c'est le promoteur qui est sujet à l'inversion. Lorsque le promoteur de H2 est correctement positionné, il y a co-transcription d'un gène codant un répresseur de H1 (**Figure 30A2**). Le complexe multiprotéique impliqué est notamment constitué d'une recombinase appelée invertase (Vanhamme et Pays, 1995). La **recombinaison réciproque** (**Figure 30B**) est plus répandue et concerne généralement des gènes près des télomères, tant pour les plasmides linéaires chez des bactéries du genre *Borrelia* que pour les nombreux chromosomes de *Pneumocystis*. La **conversion génique**, remplacement d'une copie de gène par une autre, concerne une grande diversité de gènes à localisation chromosomique variée. Nécessitant une endonucléase, ce processus a été mis en évidence pour la première fois dans le cadre du changement de « mating-type » chez *S. cerevisiae* et a été retrouvée chez *Borrelia* et *Pneumocystis* (Vanhamme et Pays, 1995, Barbour 2003, Stringer, 2002).

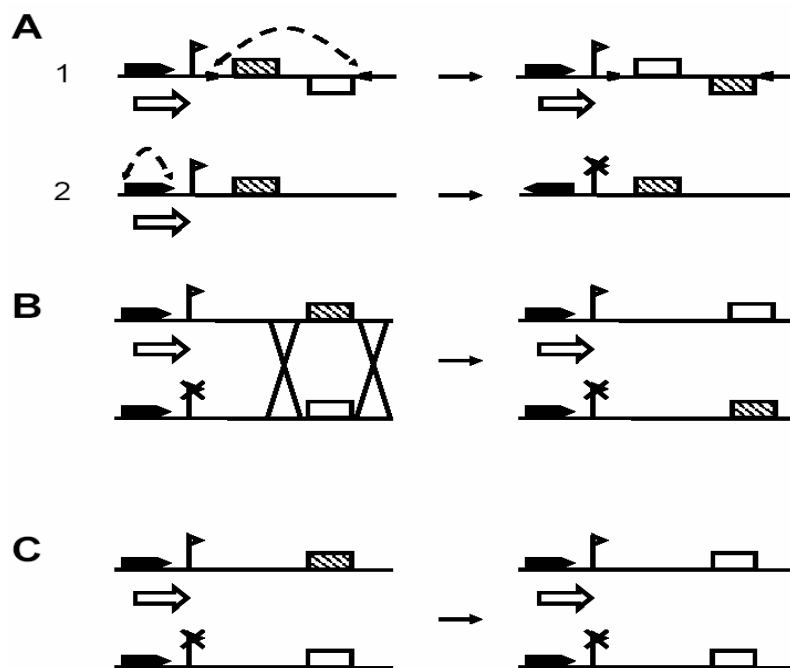


Figure 30: Représentation schématique de plusieurs modes d'activation génique par recombinaison de l'ADN. Chaque gène est représenté par un rectangle (hachurés ou non) et la direction de la transcription par la flèche blanche. Le drapeau représentant le site

d'initiation de la transcription est barré quand le promoteur (rectangle pointu noir) est inactif.

A1. Recombinaison entraînant une inversion de deux gènes à partir de courts segments d'ADN (petites flèches noires) reconnus par une ADN recombinase. **A2.** Recombinaison entraînant une inversion du promoteur. **B.** Recombinaison réciproque. **C.** Conversion génique entraînant le remplacement d'un gène qui sera perdu, par la copie d'un autre.

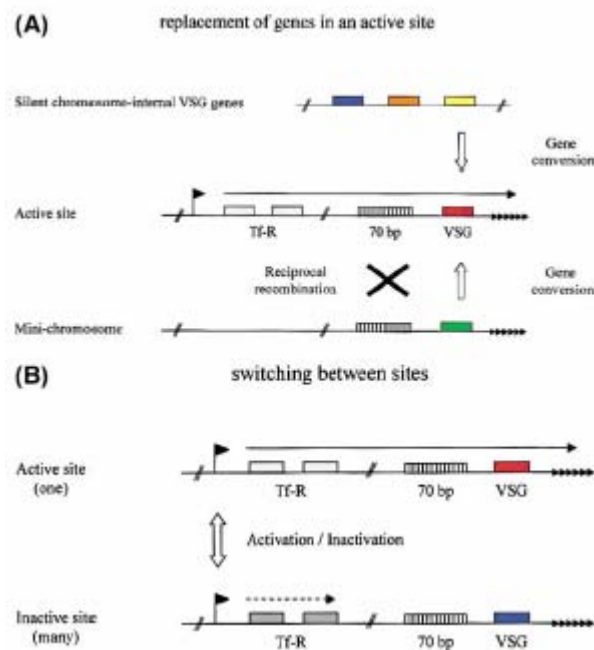


Figure 31: Mécanismes de l'exclusion allélique des gènes *vsg* chez le trypanosome. Les gènes sont représentés par des rectangles de couleurs différentes, les promoteurs par des drapeaux. **A/** Remplacement par conversion génique d'un gène par un autre au niveau d'un site actif. **B/** Remplacement d'un site actif par un autre (**Borst 2002**).

Chez *T. brucei*, le remplacement de *vsg* dans l'ES actif peut s'effectuer par recombinaison réciproque ou conversion génique (**Figure 31A**). Il est facilité par la présence de séquences répétées télomériques et conservées (70 pb) dans l'environnement du gène. L'endonucléase spécifique pour la conversion n'a pas encore été caractérisée. Il est cependant très probable que la localisation du *vsg* de l'ES actif dans une région sensible aux nucléases, correspondant à une chromatine relaxée (Pays et al, 1981), soit un facteur favorable à la conversion génique. Ce processus dépend d'homologies de séquences entre l'ADN donneur et l'ADN receveur. La localisation des blocs d'homologie peut varier considérablement entre les gènes *vsg*, d'où un site très variable de réalisation de la conversion génique (Pays et Nolan

1998). Elle peut avoir lieu en 5' ou en 3' du gène, mais elle peut aussi être indépendante des répétitions de 70 pb (McCulloch et al, 1997). Elle peut donc s'effectuer sur de très longues distances à partir de régions d'homologie éloignées ou être simplement confinée au gène lui-même, créant ainsi des gènes *vsg* hybrides. La conversion génique permet donc d'augmenter le potentiel du parasite dans la variation antigénique (Pays et Nolan 1998).

V-2-2-2 Activation *in situ*

Dans ce cas de figure, le gène est activé là où il se trouve, sans aucun déplacement de séquence d'ADN. Une variation de taille d'une simple répétition peut être utilisée soit pour activer la transcription, soit pour éviter un arrêt de la traduction. Chez les bactéries *Neisseria*, tous les gènes de la famille *opa*, codant une protéine de surface, sont transcrits de manière constitutive mais seulement certains des ARNm seront traduits. Un segment d'ARNm étant constitué d'une série de pentamères (CTCTT), l'addition ou la soustraction d'un ou plusieurs de ces pentamères survient par erreurs commises lors de la réplication d'ADN. Cela pousse la traduction hors de phase et il en résulte une synthèse de protéines tronquées (Vanhamme et Pays, 1995).

L'activation *in situ* d'un autre site d'expression, sans altération de l'ADN (Figure 31B), serait un autre moyen par lequel les trypanosomes modifient l'expression des *vsg* mais le mécanisme n'est pas encore bien compris (Pays et Nolan 1998). Selon le modèle proposé par Navarro et Gull (2001) par rapport à l'exclusion allélique, ce mécanisme s'explique par l'existence dans le noyau d'un sous-compartiment spécifique appelé **ESB** (« Expression Site Body ») où est localisé l'ES actif (**Figure 32**). L'espace procuré par l'ESB serait insuffisant pour permettre l'accès à deux ES. Il reste à savoir comment l'ES actif se retrouve dans l'ESB et comment son remplacement par un autre ES, lors de l'activation *in situ*, s'opère.

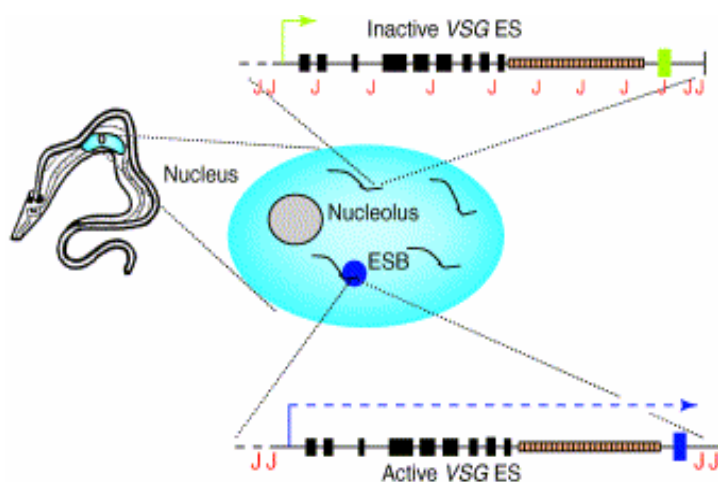


Figure 32: Schéma illustrant la notion d'activation in situ d'un gène vsg dans le noyau du trypanosome. L'ESB (en bleu) est un compartiment subnucléaire hébergeant l'ES actif (Active VSG ES). Les ES inactifs en seraient exclus. (d'après McCulloch, 2004).

L'exclusion allélique a été mise en évidence pour les gènes *var* de *P. falciparum* à l'aide de différentes approches expérimentales: RT-PCR (Chen et al, 1998), « nuclear run-on » (Scherf et al, 1998), hybridations Northern (Horrocks et al, 2004). La détection d'un seul transcrit *var* complet au stade trophozoïte en anneau montre bien une régulation au niveau transcriptionnel, mais c'est surtout par les techniques de transfection stable que l'exclusion allélique de ces gènes a été élucidée. Par ces techniques, il est possible d'évaluer l'expression des gènes *var* endogènes dans une population de parasites transformés et de voir si le promoteur modifié (utilisé pour la transfection) est reconnu comme membre de la famille *var*. La mise en place d'un pool d'oligonucléotides qui permet de mesurer simultanément les niveaux d'expression de chaque gène *var* dans le génome du parasite (souche 3D7) en utilisant la RT-PCR quantitative en temps réel (Salanti et al, 2003) a considérablement facilité les études.

Des parasites transformés avec des constructions présentant un intron *var* muté, entraînant une hyperactivation du promoteur, continuent à exprimer les gènes *var* chromosomiques endogènes ; il en est déduit que le promoteur modifié n'a pas été reconnu comme membre de la famille (Gannoun-Zaki et al, 2005). Des conclusions similaires ont été tirées d'expériences montrant que des promoteurs actifs de gènes *var*, insérés au locus subtelomérique du gène *hrp2*, n'affectaient pas l'expression des gènes *var* endogènes (revue Frank et Deitsch, 2006). Par contre, dans leurs essais de transfection avec diverses constructions, Voss et al. (2006) ont réussi à infiltrer la voie qui coordonne l'exclusion

allélique. L'activation du promoteur *var* modifié (utilisation d'un marqueur de sélection) entraîne une répression forte de l'expression de la famille entière, répression mesurée par Northern blot. Cela indique que le promoteur *var* modifié a été cette fois considéré comme membre de la famille.

V-2-3 Contrôle épigénétique

Le contrôle épigénétique de l'expression d'un gène correspond aux changements de l'état transcriptionnel d'un gène qui ne sont pas causés par des changements de facteurs de transcription ou de séquence d'ADN, mais résultent plutôt de modifications de la structure de l'ADN (par exemple, par méthylation) ou de celle des protéines chromatinienne (acétylation ou méthylation d'histones) (Henikoff et Matzke, 1997). La chromatine est en perpétuel équilibre entre états relâché et condensé suivant les effets antagonistes de multiples complexes protéiques. Le remodelage de la chromatine est donc un facteur crucial dans la régulation de la transcription (Workman et Kingston 1998).

Chez *P. falciparum*, les changements dans l'expression des gènes *var* (activation /inactivation) n'ont pas été associés à des changements de séquences autour des loci *var*, ce qui suggère que les mécanismes sous-jacents sont épigénétiques (Scherf et al, 1998 ; Deitsch et al, 1999). Les recherches sur ces mécanismes ont été menées selon trois approches différentes:

- ✓ l'analyse des changements de la structure chromatinienne et de leur effet sur la transcription.
- ✓ l'étude de la fonction des régions conservées non codantes des gènes *var*.
- ✓ la détermination de la localisation nucléaire des segments chromosomiques contenant des gènes *var*.

La régulation épigénétique de ces gènes a été confortée par les résultats d'un suivi des commutations sur 40 générations successives (Horrocks et al, 2004). Pour chaque gène *var*, il existe des taux de commutations "ON" et "OFF" intrinsèques. Considérant que les changements de structure de la chromatine locale assurent potentiellement la transition entre différents états transcriptionnels, les auteurs ont postulé que le gène *var* pouvait exister sous trois états différents: actif, inactif avec possibilité d'être activé et l'état fortement inactif (forte répression). Chez plusieurs organismes, il est connu que la structure de la chromatine est modulée par les modifications post-traductionnelles d'histones (phosphorylation, méthylation et acétylation) (Felsenfeld et al, 1996). Chez *S. cerevisiae* par exemple, il a été montré que la

protéine Sir2, une désacetylase NAD-dépendante, est impliquée dans l'inactivation (« silencing ») des loci HML et HMR par hétérochromatinisation de la région chromosomique où ces gènes sont localisés (Moazed, 2001).

Le rôle de l'homologue de Sir2 chez *P. falciparum* (**PfSir2**), a été étudié dans le contexte de la régulation de l'expression des gènes *var*. La technique d'immunoprécipitation de chromatine a révélé que PfSir2 se lie aux extrémités des chromosomes et jusqu'à 55 kb vers les centromères, c'est-à-dire des régions contenant 2/3 des copies *var*. L'activation d'un gène *var* de telles régions se traduit alors par une délocalisation de PfSir2. Par contre, les gènes *var* internes, silencieux, n'ont montré aucune liaison évidente avec PfSir2. L'inhibition de l'expression de ces derniers gènes caractérisés par un promoteur UpsC, semble donc être indépendante de PfSir2 (Freitas-Junior et al, 2005). Les régions subtélomériques de *P. falciparum* présentent bien les caractéristiques de l'hétérochromatine, et cet environnement chromatinien peut réprimer de façon réversible le promoteur d'un gène artificiellement inséré dans de telles régions (Duraisingh et al, 2005). Des mutants « knock out » du gène *PfSir2* ont été réalisés. Le fort niveau d'expression des gènes *var*, évalué sur « microarrays », confirme le rôle de PfSir2 dans l'inactivation de ces gènes. Cependant, la majorité des gènes de type UpsB et UpsC restant silencieux, le rôle de PfSir2 serait limité à la condensation de la chromatine subtélomérique et donc au contrôle de l'expression des gènes de type UpsA. Un deuxième gène Sir2-like présent dans le génome de *P. falciparum* pourrait être impliqué, ce qui augmente la complexité de cette voie de régulation (revue Frank et Deitsch, 2006).

D'autres études s'appuyant sur la technique FISH ont été menées afin de suivre l'activité des promoteurs *var* chez des souches sauvages ou génétiquement modifiées de parasites en culture. L'activation du locus subtélomérique *var2csa*, codant la molécule d'adhésion responsable de la liaison au chondroïtine sulfate A placentaire, semble être associée à une délocalisation loin du cluster télomérique (**Figure 33**) (Ralph et al, 2005). Toutefois, un promoteur central *var* activé en fusion transcriptionnelle avec le gène humain *dhfr* est resté associé au cluster (Voss et al, 2006). L'utilisation de constructions contenant des promoteurs *var* portés par des épisomes ou intégrés dans des régions subtélomériques du génome ont permis de montrer que l'activation de ces promoteurs coïncide avec leur localisation à la périphérie nucléaire (Duraisingh et al, 2005 et Voss et al, 2006). Ainsi, les promoteurs *var* transcriptionnellement actifs seraient finalement confinés dans une région subnucléaire spécifique, tout comme dans le cas de la régulation des gènes *vsg* chez le trypanosome africain. Conformément à cette hypothèse, des observations en microscopie

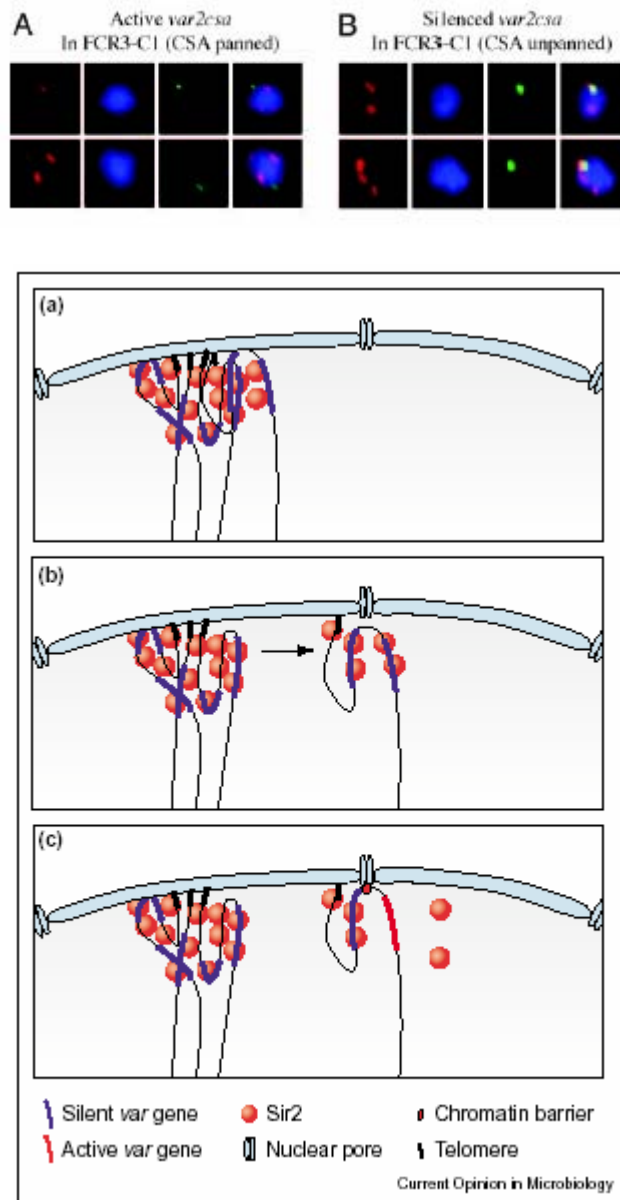


Figure 33: Localisation par la technique FISH (« Fluorescence In Situ Hybridization ») du gène *var2csa* (en vert) et de l'ADN télomérique (en rouge) dans le noyau (ADN total fluorescent en bleu après liaison au DAPI) chez *Plasmodium falciparum*. **A/** Le gène *var* transcriptionnellement actif est physiquement séparé des amas télomériques. **B/** Le même gène *var* dans un état inhibé co-localise avec un amas télomérique. **C/** Modèle d'activation génique. La migration d'un chromosome en dehors de l'amas télomérique puis la dissociation des protéines Sir2, sont nécessaires à l'expression du gène *var* (Ralph et Scherf, 2005).

électronique ont permis d'identifier une zone euchromatinienne «lâche» à l'intérieur de l'hétérochromatine de la périphérie nucléaire (**Figure 34**) (Ralph et al, 2005).

Les gènes *var* étant souvent disposés en tandem le long des chromosomes et, comme l'activation d'un gène n'est pas associée à l'activation du gène adjacent, il doit exister un autre niveau de régulation permettant de réprimer les différents gènes, même lorsqu'ils sont situés dans une région nucléaire favorable à la transcription. Frank et Deitsch (2006) ont montré que des constructions portant un promoteur *var* couplé à l'intron du gène *var* reste silencieux même si le promoteur *var* voisin est en activité. L'intron pourrait donc imposer une répression transcriptionnelle, même au sein d'une région nucléaire favorable.

L'organisation des télomères en «bouquets» est supposée importante pour la diversification des gènes *var*, une étroite relation de proximité étant normalement favorable aux recombinaisons entre gènes (Freitas-Junior et al, 2000).

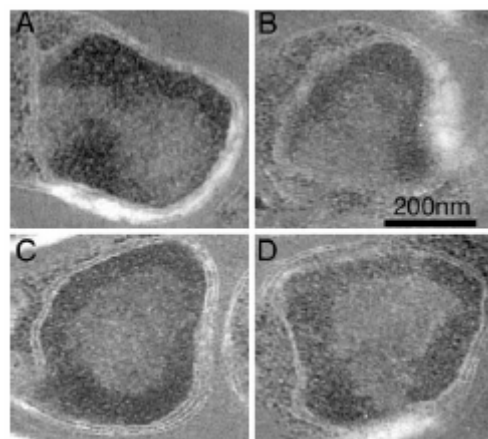


Figure 34: Mise en évidence ultrastructurale d'une zone euchromatinienne située à la périphérie du noyau, lieu majeur de l'hétérochromatine chez *P. falciparum* (Ralph et al, 2005).

V- Les microsporidies

V-1 Généralités

Le terme « **microsporidie** » est appliqué à un groupe d'Eucaryotes unicellulaires, tous parasites intracellulaires obligatoires et formant des spores équipées d'un long **tube polaire** enroulé dans la région postérieure. Depuis 1857, année de la description par Näegeli de la toute première microsporidie (*Nosema bombycis*, parasite du ver à soie), près de 1300 espèces réparties dans 143 genres ont été identifiées.

Par rapport à la plupart des autres cellules eucaryotes, l'organisation de la cellule microsporidienne est plutôt atypique, en raison de l'absence d'organites producteurs d'ATP (mitochondries ou hydrogénosomes) ainsi que de peroxyosomes. L'appareil de Golgi est difficilement repérable car n'offrant jamais l'image d'un empilement de saccules caractéristique d'un dictyosome. Selon le genre considéré, l'appareil nucléaire de la spore est limité à un seul noyau ou représenté par deux noyaux juxtaposés, constituant un diplocaryon. La phase d'infestation est marquée par le phénomène de **germination** qui implique la brusque extrusion du tube polaire à l'apex de la spore, la lumière du tube extrudé servant au transfert du sporoplasme dans une cellule-hôte potentielle. Le développement intracellulaire subséquent comprend toujours une phase de multiplication ou **mérogonie**, suivie d'une phase de différenciation ou **sporogonie**. Le mode de division cellulaire peut être simple ou multiple après formation d'un plasmode. La mitose est acentriolaire et de type fermé. La méiose n'a été mise en évidence que chez des espèces à cycle complexe faisant intervenir différents hôtes (Wittner et Weiss, 1999).

Les microsporidies infectent pratiquement tous les groupes animaux, des invertébrés aux mammifères, dans divers milieux terrestres et aquatiques. Plus de la moitié des espèces connues sont des pathogènes d'insectes, certaines ayant un impact économique notable (*Nosema bombycis* en sériciculture, *N. apis* en apiculture). La première microsporidie isolée à partir d'un hôte mammifère fut *Encephalitozoon cuniculi* (Levaditi et al. 1924), responsable d'une encéphalite granulomateuse chez le lapin (Wright et Craighead, 1922). Le premier cas avéré de microsporidiose humaine fut rapporté en 1959, pour un jeune garçon japonais, souffrant de vomissements et de maux de tête (Matsubayashi et al, 1959). Les infections chroniques ou persistantes chez les personnes immunocompétentes sont cependant souvent asymptomatiques.

Espèces microsporidiennes	Références (Première détection chez l'Homme)
<i>Brachiola</i> (syn. <i>Nosema</i>) <i>algerae</i>	Visvesvara et al. (1999)
<i>Brachiola vesicularum</i> (syn. <i>Nosema</i> -like sp.)	Cali et al. (1996)
<i>Encephalitozoon cuniculi</i>	Bergquist et al. (1984a,b)
<i>Encephalitozoon hellem</i>	Friedberg et al. (1990)
<i>Encephalitozoon</i> (syn. <i>Septata</i>) <i>intestinalis</i>	Cali et al. (1991a)
<i>Enterocytozoon bieneusi</i>	Desportes et al. (1985)
<i>Nosema ocularum</i>	Bryan et al. (1991)
<i>Pleistophora ronneafiei</i>	Ledford et al. (1985)
<i>Trachipleistophora anthropopthera</i>	Yachnis et al. (1996)
<i>Trachipleistophora hominis</i>	Field et al. (1996)
<i>Vittaforma corneae</i> (syn. <i>Nosema corneum</i>)	Davis et al. (1990)

Tableau 2: Espèces microsporidiennes infectant l'Homme.

C'est surtout chez les personnes immunodéficientes, notamment celles atteintes du SIDA, que l'incidence des microsporidies sur la santé humaine est la plus significative (**Didier, 2005**). Les espèces actuellement reconnues pathogènes de l'Homme, appartiennent aux 7 genres suivants : *Enterocytozoon*, *Encephalitozoon*, *Vittaforma*, *Pleistophora*, *Trachipleistophora*, *Nosema* et *Brachiola* (**Tableau 2**). Les infections sont souvent disséminées, parfois limitées à certains sites comme la cornée de l'œil et le muscle squelettique. La détection des mêmes espèces chez plusieurs animaux a fait envisager un potentiel zoonotique (Mathis et al, 2005). Plus ou moins efficace, la chimiothérapie des microsporidies est principalement basée sur deux types de composés : l'albendazole, se liant à la tubuline, et la fumagilline ou l'un de ses analogues (TNP-470) se liant à la méthionine-aminopeptidase II (Bacchi et Weiss, 2004).

V-2 Position phylogénétique : un changement radical

Compte-tenu de leur organisation unicellulaire, les microsporidies ont longtemps été considérées comme des Protistes. Leur rangement dans un phylum distinct, le phylum des Microspora, a été proposé par Sprague (1977). Les années 1980 furent marquées par la conception assez surprenante selon laquelle les microsporidies seraient les représentants d'une lignée eucaryote très ancienne. Selon Cavalier Smith (1998), certains groupes de protozoaires amitochondriaux (Microsporidia, Diplomonada, Archamoeba) auraient divergé avant la réalisation de l'endosymbiose primaire à l'origine des mitochondries, d'où sa proposition de les rassembler sous le vocable **Archezoa**. Une origine ancienne des microsporidies était notamment suggérée par des similitudes entre leurs ribosomes et ceux des eubactéries : mêmes constantes de sédimentation pour le ribosome entier (70S) et les deux sous-unités ribosomiques (30S et 50S) (Ishihara et Hayashi, 1968), mêmes constantes de sédimentation pour les deux ARNr majeurs (16S et 23S) (Cury et al, 1980), et même association covalente de la séquence 5,8S à l'extrémité 5' de l'ARNr 23S (Vossbrinck et Woese, 1986). Suite à une étude phylogénétique au moyen des séquences d'ARNr 16S/18S incluant celle d'une microsporidie (*Vairimorpha necatrix*), il apparut que la branche microsporidienne était ancrée près de la base de l'arbre des Eucaryotes (Vossbrinck et al, 1987). L'hypothèse de primitivité était ainsi mieux étayée.

La position basale présumée des microsporidies fut remise en question lors de travaux portant sur d'autres séquences :

- les arbres construits avec les tubulines α et β indiquaient cette fois une émergence nettement tardive, plus précisément au niveau des champignons! (Edlind et al, 1996 ; Keeling et Doolittle, 1996).

- un gène codant une HSP70 de type mitochondrial était détecté chez différentes microsporidies, suggérant que leur ancêtre possédait des mitochondries (Germot et al, 1997 ; Hirt et al, 1997 ; Peyretailade et al, 1998a).

- l'examen de la première séquence complète d'ARNr 23S microsporidien révélait une organisation typiquement eucaryotique avec délétion des parties variables et l'analyse phylogénétique prenant en compte l'hétérogénéité de la vitesse d'évolution des sites montrait une émergence tardive (Peyretailade et al, 1998b).

La majorité des études phylogénétiques ultérieures avec diverses protéines conservées (actine, ARNt-synthétases, grande sous-unité d'ARN polymérase II...) ont alors fait conclure à la relation de parenté « **Microsporidia-Fungi** ». Cette relation peut justifier quelques traits

cytologiques et biochimiques communs (production de spores, mitose fermée, présence de chitine et de tréhalose). L'évolution des gènes microsporidiens étant particulièrement rapide, il a été vérifié que le placement basal était bien dû à un artefact de reconstruction, dit «attraction des longues branches» (Van de Peer et al, 2000, Thomarat et al, 2004). Il faut aussi noter que l'évolution rapide fait obstacle à l'établissement d'une connexion suffisamment robuste pour préciser le groupe fongique le plus proche des microsporidies (Zygomycota, Chytridiomycota...?). Une phylogénie des Fungi portant sur 199 espèces a récemment été déduite de l'analyse combinée de 6 gènes différents (James et al, 2006). Sa branche la plus précoce est représentée par les microsporidies (*E. cuniculi* et *Antonospora locustae*) groupées avec le chytridiomycète *Rozella allomyces* qui est un parasite intracellulaire d'autres champignons et qui, semblablement aux microsporidies, perd sa paroi quand il entre dans une cellule-hôte. Bien qu'ayant en commun la capacité de former des spores flagellées (zoospores), les chytridiomycètes ne constitueraient pas un groupe monophylétique. Leur répartition dans l'arbre a alors amené les auteurs à considérer que l'évolution fongique a impliqué au moins quatre pertes indépendantes du flagelle. Dans cette hypothèse, les formes ancestrales flagellées vivaient exclusivement en milieu aquatique et la perte de locomotion flagellaire chez certains de leurs descendants aurait été compensée par de nouveaux modes de sporogenèse.

La présence d'un organite cryptique dérivé de la mitochondrie (« mitosome ») chez les microsporidies a été postulée suite à l'annotation de la séquence génomique d'*E. cuniculi* révélant des homologues de différentes protéines mitochondriales ayant pour fonction principale la biosynthèse des groupes Fe-S (Katinka et al, 2001). Une preuve expérimentale a été apportée par une étude immunocytochimique montrant la localisation de l'orthologue d'HSP70 mitochondriale au niveau de petites vésicules délimitées par deux membranes chez *Trachipleistophora hominis* (Williams et al, 2002). Certains composants du système d'adressage mitochondrial sont également préservés pour ces mitosomes, mais le mécanisme d'importation serait plus simplifié chez *E. cuniculi* que chez *A. locustae* (Burri et al, 2006).

V-3 Structure de la spore

Chez de nombreuses microsporidies, la spore est ovoïde ou pyriforme, avec une longueur souvent située entre 1 et 5 µm (1-1,5 µm chez *Enterocytozoon bieneusi*, 2-3 µm chez *Encephalitozoon* spp). Certaines espèces sont cependant caractérisées par des spores de

grande taille, en forme de bâtonnets plus ou moins longs (jusqu'à 40µm chez *Bacillidium filiferum*).

La spore est entourée d'une paroi chitino-protéique très résistante, essentielle à la survie du parasite dans l'environnement. Cette paroi est classiquement subdivisée en deux couches : l'**exospore**, désignant une fine couche externe opaque aux électrons, riche en glycoprotéines, et l'**endospore**, couche interne plus épaisse intimement accolée à la membrane plasmique et transparente aux électrons du fait de sa grande richesse en chitine. Elle protège efficacement la cellule sporale des agressions externes mais elle contribue aussi, par sa rigidité, au maintien d'une forte pression hydrostatique interne qui sera utilisée pour l'extrusion du tube polaire pendant la germination (Frixione et al, 1997).

L'intérieur de la spore comprend l'appareil nucléaire (un seul noyau ou un diplocaryon) et un cytoplasme compact, riche en ribosomes, largement occupé par un ensemble de structures constituant l'appareil invasif ou appareil d'extrusion (Franzen C. et Müller A. 1999 ; Bigliardi E. et Sacchi L. 2001). L'élément principal de cet appareil est le **tube polaire**, enroulé en spirale dans la région postérieure (de 4 à 30 tours de spire, selon l'espèce). Le segment antérieur et rectiligne du tube polaire (manubrium) rejoint une structure apicale en forme de dôme, dite **capuchon polaire**. Bien que parfois pris comme synonyme de « capuchon polaire », le terme « disque d'ancrage » désigne normalement un territoire spécifiquement accolé à l'extrémité du tube polaire, la région supérieure plus étalée étant le « sac polaire ». Un autre élément de l'appareil invasif est le **polaroplaste**, un système membranaire encerclant le manubrium et apparaissant comme un empilement de saccules très aplatis (polaroplaste lamellaire) et de vésicules au contenu dense (polaroplaste vésiculaire, plus proche du noyau). L'espace délimité par la partie enroulée du tube polaire est généralement occupé par une grande vacuole riche en éléments opaques aux électrons, la **vacuole postérieure** (Figure 35).

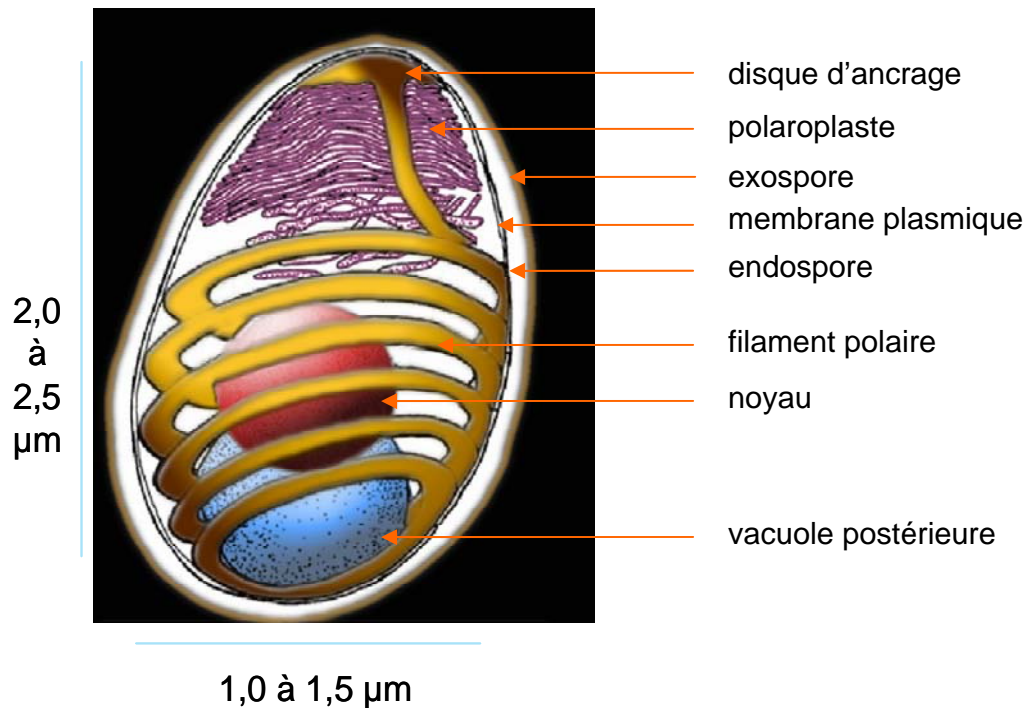


Figure 35: Structure classique d'une spore microsporidienne (Franzen et Müller, 1999).

V-4 Germination et invasion cellulaire

La germination d'une spore microsporidienne a pour but l'inoculation du sporoplasme à une cellule-cible. Elle commence par l'extrusion extrêmement rapide (quelques millisecondes) du tube polaire au pôle antérieur de la spore (Frixione et al, 1992). Le tube extrudé s'étire sur plusieurs dizaines de µm tout en restant attaché à la spore par l'une de ses extrémités. Observable en microscopie à contraste de phase sur des spores purifiées, ce phénomène est inductible *in vitro* sous l'influence de divers stimuli avec une efficacité variable d'un genre à l'autre (dessiccation suivie de réhydratation ; variations de pH ou d'osmolarité ; addition de cations monovalents, d'urée, de peroxyde d'hydrogène, de mucine...). Le sporoplasme, matériel intrasporal représentatif de l'entité infectieuse, transite ensuite dans la lumière du tube et forme une gouttelette visible à l'extrémité libre du tube. Un suivi en vidéo-microscopie a fait estimer à 2 secondes la durée totale des événements conduisant à la décharge complète du sporoplasme (Frixione et al, 1992). La nature des stimuli physiologiques reste sujette à spéculation. L'examen du sporoplasme en microscopie électronique révèle une cellule très simplifiée, limitée par la seule membrane plasmique, les structures internes les plus évidentes étant le(s) noyau(x), les ribosomes et quelques vésicules.

La spore vide conserve sa membrane. La nouvelle membrane affectée au sporoplasme tire vraisemblablement son origine du polaroplasme lamellaire.

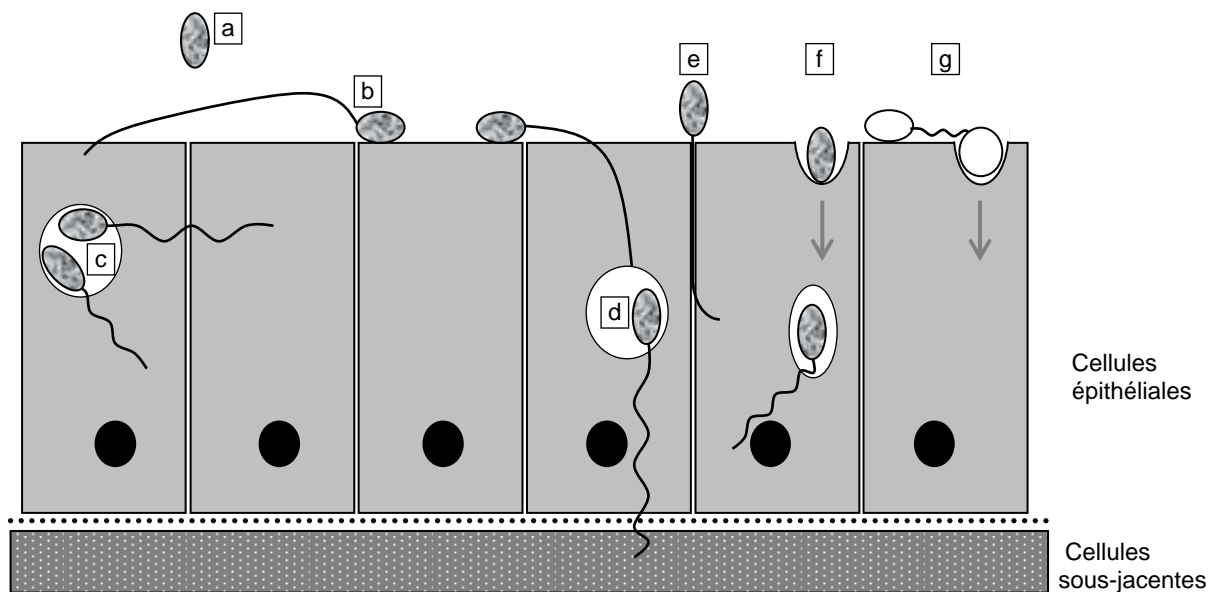


Figure 36: Les différentes hypothèses sur les voies d'entrée ((b à g) d'*Encephalitozoon* spp. dans les cellules-hôtes (d'après **Vivares et Méténier, 2001**).

(a) Spore libre. (b) La spore adhère à la surface de la cellule-hôte. Le tube polaire est expulsé et perce la membrane plasmique de cette dernière pour inoculer le sporoplasme. (c) Infestation secondaire de la même cellule-hôte ou d'une cellule-hôte adjacente, après prolifération du parasite dans une vacuole parasitophore. (d) Infestation secondaire d'une cellule sous-jacente (autre tissu ?). (e) Passage du tube polaire extrudé via l'espace intercellulaire puis pénétration latérale. (f) Endocytose de la spore entière précédant la germination. (g) Endocytose du seul sporoplasme après germination.

Le mécanisme par lequel le sporoplasme pénètre à l'intérieur d'une cellule-hôte reste très incertain. L'hypothèse usuelle est que la décharge du tube polaire provoquerait une rupture mécanique locale de la membrane plasmique de la cellule-hôte, conduisant directement à l'injection du sporoplasme dans le cytoplasme de celle-ci (Weidner et al, 1984). Cependant, certaines observations après infestation de cellules en culture indiquent que la spore microsporidienne peut être internalisée par phagocytose. L'extrusion secondaire du tube polaire permettrait alors au parasite d'expulser le sporoplasme en dehors du phagosome (Franzen, 2005) (**Figure 36**). L'origine de la membrane de la vacuole parasitophore formée chez *Encephalitozoon* spp est problématique.

V-5 Développement intracellulaire

Au terme de la phase d'invasion, le sporoplasme entre dans une phase de prolifération active dite **mérogonie**. Chez certains genres, par exemple *Encephalitozoon*, *Nosema*, *Vittaforma* ou *Brachiola*, la division de l'appareil nucléaire (monocaryon ou diplocaryon) est immédiatement suivie de celle du cytoplasme (fission binaire), produisant des cellules individualisées, appelées mérontes. Ces cellules au cytoplasme peu différencié ont une localisation souvent périphérique (Franzen et Müller, 1999). Chez d'autres genres comme *Enterocytozoon*, *Pleistophora*, *Trachipleistophora*, un plasmode mérogonial est formé, l'appareil nucléaire subissant une série de mitoses en l'absence de cytodiérèse (Cali et Takvorian, 1999). La division nucléaire se déroule pour la plupart par pleuromitose (Figure 37), sans centrioles mais avec la formation de plaques polaires (ou SPB pour « spindle pole bodies »), similaires à celles de levures (Vivarès et al, 1977). Ce type de mitose est caractérisé par des hémi-fuseaux indépendants à l'intérieur du noyau, une faible condensation des chromosomes et l'absence d'arrangement de ces derniers en plaque équatoriale. Au début, le noyau n'est pourvu que d'une seule plaque polaire attachée à un seul fuseau intranucléaire avec des microtubules liés aux kinétochores des chromosomes. Cette plaque polaire est ensuite dupliquée et deux hémi-fuseaux se séparent. En fin de mitose, les plaques atteignent les pôles opposés du noyau et les fuseaux s'étirent pour aboutir à la séparation des deux jeux de chromatides (Raikov, 1994).

Les mérontes ou le plasmode mérogonial s'engagent ensuite dans la phase de **sporogonie**. La commutation en sporontes ou plasmode sporogonial est marquée par le dépôt d'un matériel de surface très opaque aux électrons, un accroissement du nombre de ribosomes et un développement plus important du réticulum endoplasmique (Cali et Takvorian, 1999). Les divisions subséquentes produisent des sporoblastes qui se transforment progressivement en spores, cette différenciation impliquant la biogenèse de l'appareil invasif et de la paroi bipartite ainsi qu'une importante compaction cytoplasmique. Les spores matures peuvent être libérées dans l'environnement après lyse de la cellule hôte (**Figure 38**).

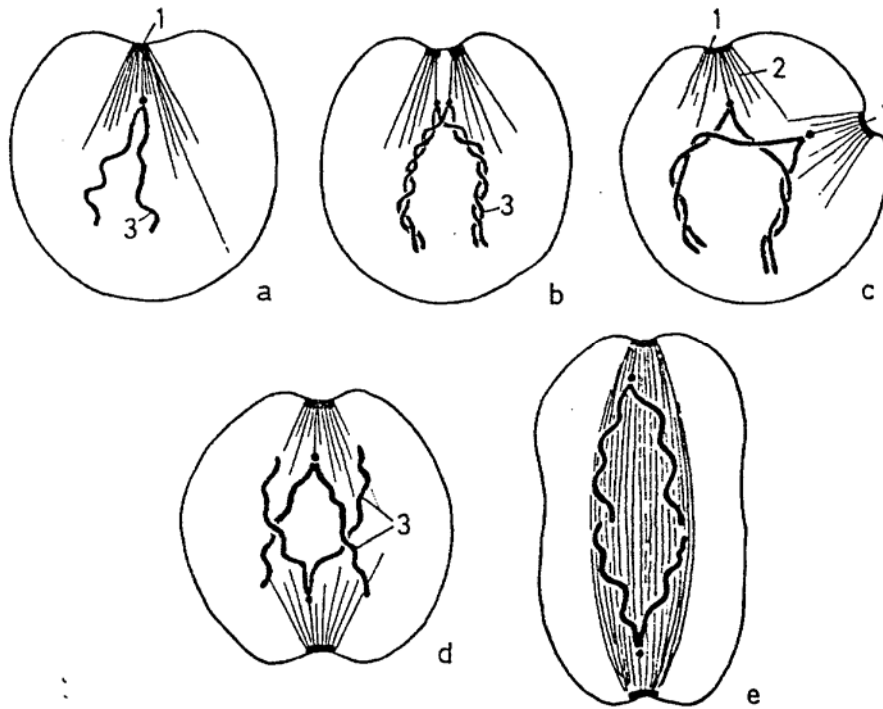


Figure 37: Stades d'une pleuromitose intranucléaire fermée (d'après **Raikov, 1994**).

a. Interphase; **b, c.** Prophase; **d, e.** Anaphase. **1** = Centre organisateur de microtubules (plaque polaire), **2** = Héli-fuseau mitotique, **3** = Chromosome (pour plus de clarté, un seul chromosome avec son kinétochore médian est représenté dans ce schéma).

Le développement des microsporidies s'effectue très souvent au contact direct du cytoplasme de la cellule-hôte. C'est le cas chez *Enterocytozoon*, *Nosema* et *Brachiola*. Le réticulum endoplasmique de la cellule-hôte peut se réorganiser pour entourer les mérontes (*Vittaforma*...). Une interface hôte-parasite plus évidente est la « vésicule sporophore », qui correspond à une sécrétion s'accumulant à la périphérie des parasites pendant la sporogonie (*Pleistophora*, *Trachipleistophora*...). Plus rarement, une vacuole parasitophore délimitée par une membrane spécifique assure le confinement de tous les stades parasitaires (*Encephalitozoon*) (**Figure 39**). Les mérontes sont intimement accolés à la membrane vacuolaire. La conversion en sporontes s'accompagne de la dissociation de cette jonction, si bien que les stades suivants prennent une position de plus en plus centrale.

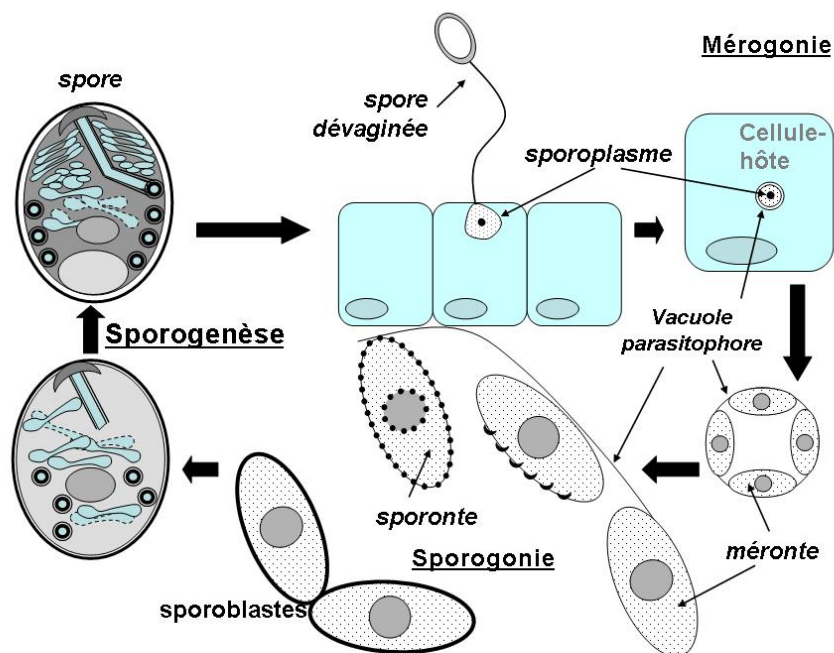


Figure 38 : Schéma du cycle de reproduction d'*Encephalitozoon cuniculi*. Après entrée du sporoplasme dans la cellule-hôte, tout le développement de cette microsporidie se déroule à l'intérieur d'une vacuole parasitophore.

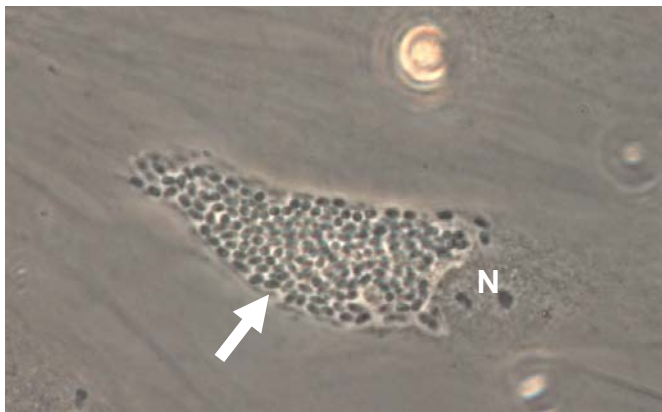


Figure 39 : Cellule MDCK infestée par *Encephalitozoon cuniculi*, observée en microscopie optique. La flèche montre une grande vacuole parasitophore contenant de nombreux parasites. N, noyau de la cellule-hôte.

Certaines microsporidies, notamment les parasites de crustacés et de poissons, peuvent induire l'hypertrophie de la cellule-hôte. La forte accumulation des stades parasitaires dans

une même cellule-hôte hypertrophiée contribue à l'élaboration d'une structure de plus ou moins grande taille, parfois visible à l'œil nu. Ce complexe morpho-physiologique, appelé xénome, est maintenu sur de longues périodes sans entraîner la mort de l'hôte. Son organisation serait plus dépendante des propriétés du parasite que de celles de l'hôte (Lom et Nilsen, 2003).

V-6 Diversité des cycles biologiques

Les microsporidies affichent une grande hétérogénéité aussi bien par rapport au nombre d'hôtes nécessaires pour assurer le cycle complet qu'à la spécificité d'hôte. Ceci est parfaitement illustré par des espèces parasites d'insectes.

Des espèces comme *Brachiola algerae* et *Vavraia culicis* présentent un cycle simple avec un seul hôte (cycle monoxène) mais aussi une seule forme sporale transmise par voie horizontale. Ces deux organismes se singularisent également par une faible spécificité d'hôte, et même du tissu infesté (Becnel et al, 2005). *B. algerae* est capable de se multiplier dans une large gamme de températures (de 24°C à 38°C), ce qui explique en partie ses capacités d'infestation et son large éventail d'hôtes potentiels (Lowman et al, 2000). De récentes études ont montré des cas d'infection de l'Homme par cette microsporidie, mais les mécanismes de la transmission restent encore mal définis (Coyle et al, 2003)

Les genres *Amblyopora* et *Edhazardia* présentent un cycle complexe qui implique deux générations d'hôtes, avec une phase sexuée (gamétogenèse, plasmogamie) qui se déroule chez la larve d'insecte (**Figure 40**). Ces microsporidies sont polymorphes (existence de 3 formes sporales au cours du cycle) avec une transmission par voies horizontale et verticale (Becnel et al, 2005). Chez *Amblyospora*, le cycle passe nécessairement par un hôte intermédiaire copépode (Andreadis, 2005). Même si certains d'entre eux peuvent utiliser différents copépodes comme hôte intermédiaire, les Amblyosporidae font montre d'une grande spécificité par rapport à leur hôte insecte définitif, contrairement à *Brachiola* et *Vavraia* (Vossbrinck et al, 2004). Hazard et Brookbank (1984), ont bien décrit en microscopie photonique la caryogamie et les différentes étapes de la méiose chez *Amblyospora sp.* Des observations en microscopie électronique ont révélé chez quelques autres genres microsporidiens la présence de complexes synaptonémaux, indicateurs de la division réductionnelle lors de la méiose (Loubès, 1979, Larsson, 1986b ; Vivarès et Sprague 1979).

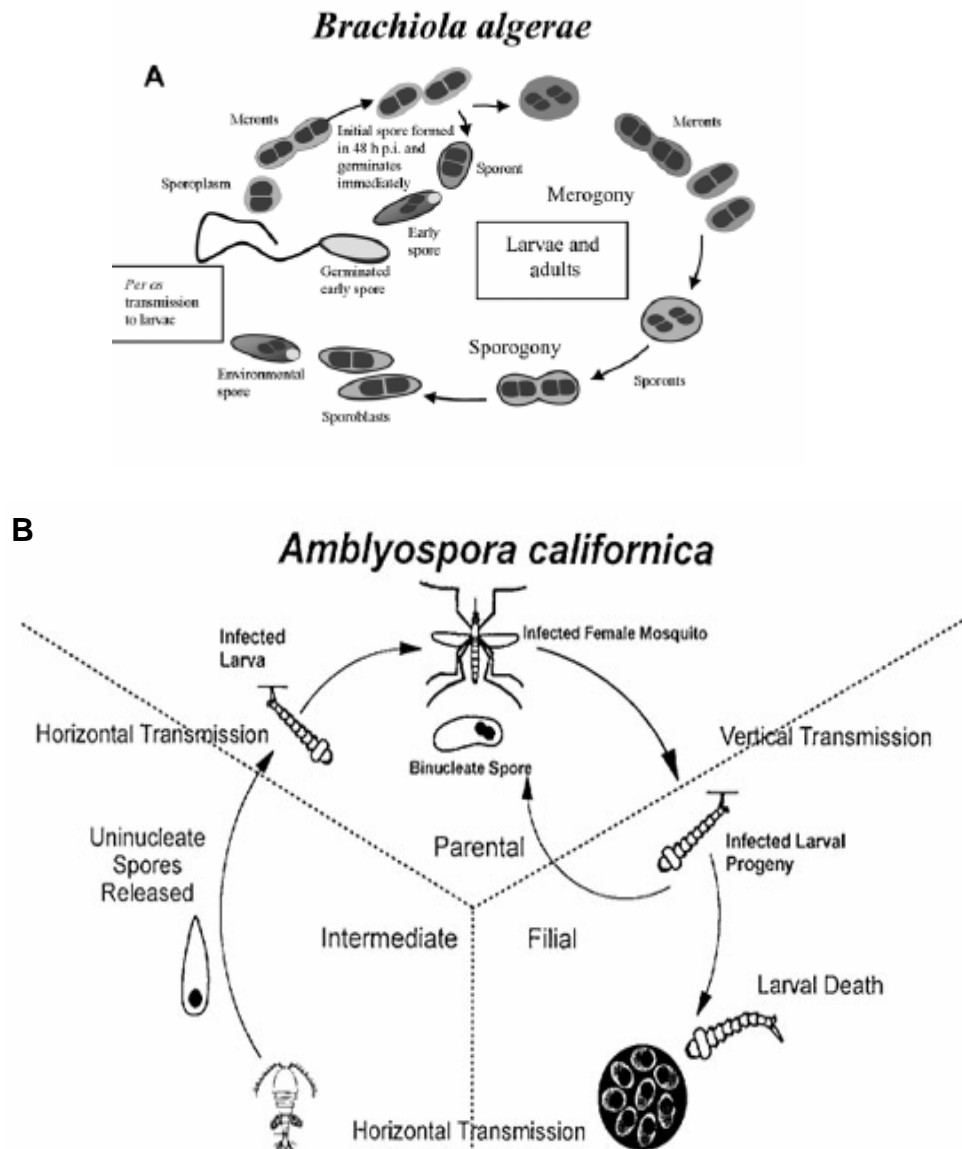


Figure 40 : Cycles de développement de deux microsporidies parasites de moustiques, *Brachiola algerae* (A) et *Amblyospora californica* (B).

Un cas intermédiaire concerne le genre diplocaryotique *Nosema* : le cycle est monoxène mais s'accompagne de dimorphisme sporal (Vavra et Larsson, 1999). Chez *N. bombycis*, la première phase de sporulation produit des spores caractérisées par une paroi peu épaisse et un filament polaire présentant seulement 4 tours de spire, d'où le nom « FC (Few Coils of polar tube) spore ». Ce sont les spores primaires. Celles-ci peuvent ré-infester d'autres cellules du ver à soie (dans les gonades, par exemple) et peuvent être transmises verticalement (transmission dite "transovarienne"). La seconde phase de sporulation donne des spores avec

une paroi plus épaisse et un long filament polaire formant 14 tours de spire (Iwano et al, 1991).

V-7 Génome nucléaire

Le matériel génétique des microsporidies est exclusivement localisé dans l'appareil nucléaire. Les espèces présentant toujours un seul noyau sont dites monocaryotiques. Les autres, plus nombreuses, sont diplocaryotiques mais il convient de distinguer : (1) les espèces « diplocaryotiques monomorphes » qui conservent le diplocaryon durant tout leur cycle biologique, (2) les espèces « diplocaryotiques polymorphes » qui montrent une alternance de stades uni- et binucléés lors d'un cycle complexe de type *Amblyospora*. Les deux noyaux accolés composant le diplocaryon se divisent de façon égale et synchrone. Le niveau de ploïdie a rarement été défini.

V-7-1 Caryotype moléculaire

La rareté des observations sur la méiose microsporidienne a été telle que le caryotype n'a été décrit par voie cytologique que chez une espèce d'*Amblyospora*, pour laquelle il a été dénombré 7 chromosomes (Hazard et al, 1979). Ce sont essentiellement les techniques d'électrophorèse en champs pulsés qui ont permis d'apporter des informations sur le nombre et la taille des molécules d'ADN chromosomique puis d'évaluer la taille du génome haploïde.

V-7-1-1. Variations interspécifiques

Les données sur les caryotypes moléculaires de différentes microsporidies sont résumées à la **Table 3**. Le nombre de bandes chromosomiques fluctue entre 8 et 18, avec des tailles variant entre 2,3 et 19,5 Mb. Le fait marquant est une tendance vers une réduction considérable de la taille du génome entier.

Les plus petits génomes microsporidiens connus à ce jour sont ceux du genre *Encephalitozoon*: 2,3-2,4 Mb chez *E. intestinalis* (Peyretailade et al, 1998 ; Sobottka et al, 1999), 2,4-2,5 Mb chez *E. hellem* (Sobottka et al, 1999 ; Delarbre et al, 2001) et 2,9 Mb chez *E. cuniculi* (Biderre et al, 1995). Nettement inférieures à celle du chromosome d'*Escherichia coli* K12 (4,6 Mb ; Blattner et al, 1997), ces tailles chevauchent la gamme offerte par d'autres bactéries telles que celles du genre *Staphylococcus* (de 1,8 Mb chez *S. thermophilus* à 2,9 Mb chez *S. aureus* ; <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1>) .

ESPECE MICROSPORIDIENNE	TAILLE GENOMIQUE (Mb)	NOMBRE de CHROMOSOMES	Références
<i>Glugea atherinae</i>	19,5	16	Biderre et al, 1994
<i>Nosema bombycis</i>	15,3	18	Kawakami et al, 1994
<i>Nosema pyrausta</i>	10,5	13	Munderloh et al, 1990
<i>Nosema furnacalis</i>	10,2	13	Munderloh et al, 1990
<i>Vairimorpha sp.</i>	10,2	8	Malone et McIvor, 1993
<i>Vavraia oncoperae</i>			
chenille	10,2	16	Malone et McIvor, 1993
ver	8,0	14	Malone et McIvor, 1993
<i>Nosema costelytrae</i>	7,4	8	Malone et McIvor, 1993
<i>Nadelspora canceri</i>	7,4	10	Amogan et al, 2006
<i>Spraguea lophii</i>	6,8	15	Mansour et al, 2004
<i>Nosema locustae</i>	5,3	18	Streett, 1994
<i>Paranosema grylli</i>	4,5	17	Nassonova et al, 2005
<i>Encephalitozoon cuniculi</i>	2,9	11	Biderre et al, 1995
<i>Encephalitozoon hellem</i>	2,6	13	Biderre et al, 1999a
<i>Encephalitozoon intestinalis</i>	2,3	10	Biderre et al, 1999a

Tableau 3 : Taille du génome haploïde et nombre de chromosomes chez quelques microsporidies (évaluations d'après les caryotypes moléculaires).

La taille maximale atteint 19,5 Mb pour *Glugea atherinae* (Biderre et al, 1994), une microsporidie de poisson formant des spores nettement plus grandes (~ 8 µm) que celles d'*Encephalitozoon* spp (~ 2 µm). Ainsi, le rapport entre les valeurs C minimale et maximale chez les microsporidies est actuellement de 8,4. La variation au sein d'un genre donné semble particulièrement importante pour *Nosema* : de 7,4 Mb (8 chromosomes) chez *N. costelytrae* à 15,3 Mb (18 chromosomes) chez *N. bombycis* (Munderloh et al, 1990, Malone et McIvor, 1993 ; Kawakami et al, 1994).

Notons que des spores monocaryotiques et diplocaryotiques co-existent dans les xénomes induits par *Spraguea lophii*, ce qui faisait parfois douter de l'homogénéité de cette espèce. Les caryotypes moléculaires des deux types de spores isolées à partir d'un même

xénome se sont révélés identiques, suggérant la conservation du même lot de chromosomes pour une seule espèce (Mansour et al, 2004).

V-7-1-2 Variations intraspécifiques

Une variabilité intraspécifique de la taille des chromosomes a été signalée pour la première fois chez *Vavraia oncoperae*, après comparaison de deux isolats issus de deux hôtes insectes différents. L'incidence sur la valeur haploïde est plutôt notable : 8 Mb *versus* 10,2 Mb (Malone et McIvor, 1993).

Chez *E. cuniculi*, trois souches (I, II et III) aux profils antigéniques distincts ont en effet été clairement différenciées au niveau du génotype sur la base de la variation du nombre de répétitions d'un tétranucléotide (GTTT) dans l'espaceur transcrit interne (ITS) de l'ADNr (Didier et al, 1995). Six variants caryotypiques (A-F) ayant été distingués parmi 15 isolats différents, le polymorphisme chromosomique a pu être mise en évidence pour les souches I et II (Biderre et al, 1999). La comparaison de certains de ces variants au moyen d'une technique d'analyse bidimensionnelle (KARD) a permis de préciser ce polymorphisme au niveau des fragments de restriction de plusieurs chromosomes (Brugère et al, 2001). Les événements d'insertion-délétion (Indels) couvrent chacun 3 à 5 kb et concernent principalement les extrémités chromosomiques. La diploïdie du noyau d'*E. cuniculi* est bien perçue lorsque le polymorphisme affecte certains chromosomes homologues. Chez *E. hellem*, trois génotypes différant par l'ITS de l'ADNr peuvent être aussi distingués (Mathis et al, 1999) et le caryotype électrophorétique est également sujet à variation (Sobottka et al, 1999 ; Delarbre et al, 2001).

La plasticité du génome semble particulièrement grande chez *Paranosema grylli*, espèce diplocaryotique monomorphe parasitant le grillon (Nassonova et al, 2005). Le profil PFGE unidimensionnel des spores isolées à partir du corps gras de chaque insecte est caractérisé par une trentaine de bandes migrant entre 135 et 485 kb et d'intensité très variable. Les résultats d'analyse KARD ont fait conclure à un lot de 17 chromosomes dont les homologues sont extrêmement polymorphes, cette hétérogénéité étant mise en relation d'une part avec la non-clonalité de la population parasitaire et d'autre part avec l'évolution différentielle des deux noyaux haploïdes du diplocaryon.

V-7-1-2 Cartographie en 2D (par la technique KARD-PFGE)

La technique KARD-PFGE (« Karyotype And Restriction Display ») est une technique de 2D PFGE qui assure la résolution rapide des fragments de restriction d'un génome, pour une enzyme donnée, chromosome par chromosome, sous réserve que ce génome soit fractionnable par PFGE et que le nombre de sites de restriction ne soit pas trop élevé. Diverses analyses sur le génome entier à une échelle subcaryotypique peuvent alors être envisagées.

Dans le cas d'un organisme où aucune carte physique n'est disponible, cet outil permet la discrimination d'isolats sur des bases moléculaires. Des chromosomes homologues dont la taille ne varie que de 3 à 5 kb ont pu être discriminés sans hybridation moléculaire (Brugère et al, 2000a ; Brugère et al, 2000c), montrant ainsi l'intérêt de cette technique dans la détermination de la ploïdie. De plus, des éléments caractéristiques du génome sont directement mis en évidence, comme par exemple la zone de répétition en tandem des gènes ADNr chez *S. cerevisiae* (Brugère et al, 2000b) ou la structure particulière des régions télomériques chez *E. cuniculi* (Brugère et al, 2000a). Des modifications chromosomiques doivent être détectables, comme les translocations et les cassures-réparations de chromosomes qui sont particulièrement importantes en physiopathologie des parasites (Lanzer et al, 1995).

V-7-2 Données du séquençage systématique

V-7-2-1 Chez *Encephalitozoon cuniculi*

L'annotation des séquences des 11 chromosomes composant le génome d'*E. cuniculi* a révélé près de 2000 gènes potentiels codant les protéines (Katinka et al, 2001). La densité génique est très élevée, soit un gène par kb pour l'ensemble des « cores » chromosomiques. A ce jour, le seul génome nucléaire présentant une densité génique similaire est un minuscule génome (550 kb réparti sur 3 chromosomes) appartenant au noyau résiduel (nucléomorphe) d'un endosymbionte algal hébergé par la cryptomonadine *Guillardia theta* (Douglas et al, 2001). Cette forte densité est corrélable à la réduction des espaces intergéniques (129 pb en moyenne chez *E. cuniculi*) mais aussi à la compaction des gènes eux-mêmes. Seulement une quinzaine d'introns très courts (moins de 60 pb) ont été détectés, principalement au début de la région 5' des gènes de certaines protéines ribosomales. De la comparaison avec les homologues d'autres organismes, il ressort que les séquences traduites sont aussi sujettes au

raccourcissement : 80% des protéines sont plus courtes que les homologues chez *S. cerevisiae*, le taux moyen de réduction atteignant 15% (Katinka et al, 2001). Les alignements multiples de séquences montrent des délétions dans les régions non essentielles des protéines, l'étendue des délétions étant particulièrement évidente dans de grandes ATPases telles que la dynéine et la midasine (Vivarès et al. 2002). L'assignation des domaines protéiques conservés de la banque de données CDD (NCBI) a confirmé que le protéome potentiel d'*E. cuniculi* présentait l'organisation en domaines la plus simple parmi tous les Eucaryotes unicellulaires jusqu'alors étudiés (Koonin et al, 2004). Les duplications géniques sont peu nombreuses et aucun élément transposable n'a été mis en évidence. Par contre, de courtes séquences répétées ont été identifiées dans les régions subterminales des différents chromosomes. L'assemblage quasi-complet des extrémités du chromosome I (Peyret et al, 2001) a révélé une seule copie d'ADNr 16S-23S au niveau de chaque subtélomère, ceci en accord avec les données de cartographie par 2D-PFGE concluant à un total de 22 copies pour les 11 chromosomes (Brugère et al, 2000).

En dépit de la compaction génique, le potentiel de codage du génome d'*E. cuniculi* n'est pas très grand, ce qui suppose une substantielle perte de gènes au cours de l'évolution du parasite. Nombre de déficits sont corrélables à la forte dépendance nutritionnelle et énergétique vis-à-vis de la cellule-hôte. Les gènes pour des enzymes impliquées dans la biosynthèse de petites molécules (acides aminés, oses, acides gras) sont effectivement très fortement sous-représentés. Le métabolisme des nucléotides apparaît limité aux principales interconversions. La grande majorité des protéines mitochondriales ne sont pas codées, tant pour la production d'énergie (cycle de Krebs, β -oxydation des acides gras, chaîne respiratoire et phosphorylation oxydative) que pour la réplication et l'expression d'un génome extranucléaire. L'éventail pour la plus grande famille de transporteurs (famille ABC) s'est révélé restreint à 13 membres (Cornillot et al, 2002). Certains complexes multi-protéiques conservés sont apparemment simplifiés. En fait, la diversité des familles et superfamilles protéiques connues se trouve considérablement réduite, même par rapport à d'autres parasites. Seulement 364 super-familles ont été dénombrées chez *E. cuniculi*, contre 620 chez *Candida glabrata* et 566 chez *Plasmodium falciparum* (Madera et al. 2004, <http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY>). Toutefois, près de la moitié des protéines prédites étant à fonction inconnue et l'évolution des gènes microsporidiens étant très rapide, il ne peut être exclu que la divergence de certaines séquences communes moins conservées (sous-unités régulatrices, par exemple) est devenue telle qu'elle ne permet plus de déceler des homologies significatives.

V-7-2-2 Chez *Antonospora locustae*

La taille du génome haploïde de la microsporidie d'insecte *Antonospora locustae* (ex-*Nosema locustae*) a été estimée après PFGE à 5,3 Mb (Street, 1994). Un séquençage partiel (685 kb) de ce génome a permis un début de comparaison avec le génome d'*E. cuniculi*, (Slamovits et al, 2004). Bien que les deux espèces considérées soient phylogénétiquement éloignées, le degré de synténie est élevé. Le pourcentage de gènes dans le même contexte atteint 13% et la succession d'autres gènes (30%) ne diffère que par quelques réarrangements. Contrastant avec la rapidité d'évolution des séquences, cette apparente stabilité de l'ordre des gènes dans des génomes microsporidiens a alors été supposée en relation avec la compaction et la faible fréquence de recombinaison. Le projet de séquençage du génome entier d'*A. locustae* a actuellement produit un ensemble de contigs portant 2606 ORF (<http://gmod.mbl.edu/perl/site/antonospora01?page=orfs>).

V-7-2-3 Chez *Vittaforma corneae*

Quelques éléments d'informations sur le génome de *Vittaforma corneae*, microsporidie responsable d'une kératite de la cornée chez l'Homme, ont été apportés après séquençage aléatoire de fragments représentant un total de 41 kb (Mittleider et al, 2002). Les séquences uniques couvrent 26,7 kb et concernent 21 gènes différents. L'usage des codons semble très différent de celui d'*E. cuniculi*. La détection d'un homologue de transcriptase inverse plaide en faveur de la présence de certains rétrotransposons.

Résultats

Chapitre 1: Régionalisation et organisation mosaïque des subtelomères des chromosomes de la microsporidie *Encephalitozoon cuniculi*.

Subtelomere organization in a microsporidian parasite (*Encephalitozoon cuniculi*).

Article 1:

Lavie Laurence*, Faye Ngor* Dia Ndongo, Yeramian E., Duroure C., Méténier Guy, Toguebaye Bhen S, Vivares Christian and Cornillot Emmanuel

Ce premier article présente une étude sur l'organisation des extrémités des 11 chromosomes de la microsporidie *Encephalitozoon cuniculi*. Un travail de cartographie physique (préliminaire au séquençage) avait révélé l'association d'une unité ADNr 16S-23S à chaque extrémité chromosomique. Lors du séquençage du génome, seulement trois unités ADNr ont été intégrées aux séquences consensus des chromosomes (deux pour le chromosome I et une pour le chromosome IV), illustrant nettement le caractère incomplet de l'assemblage des régions subtelomériques et télomériques.

La première étape de ce travail a été de rechercher et comparer les séquences répétées des consensus chromosomiques, à l'aide du logiciel Miropeat. Des segments homologues (sans « gaps »), de plus de 1 kb, ont ainsi été détectés dans différents chromosomes au niveau des extrémités. Quinze types de répétitions (r01 à r15) sont distingués, la répétition r01 étant caractérisée par une unité ADNr. La conservation de séquence sur de plus longues distances (avec la présence de quelques insertions/délétions) a conduit à définir 8 blocs de répétitions (EXT1-EXT8). Par exemple, le bloc EXT1 sur chaque extrémité du chromosome I comprend la série r02-r03-r04-r15, en allant de l'ADNr vers le cœur du chromosome, et les variants présumés de EXT1 sur d'autres extrémités chromosomiques sont réduits à r02-r03-r04, r02-r03 ou seulement r02. Les blocs EXT partagent entre eux des régions d'homologies correspondant à des CDS prédits par l'annotation. Ces CDS forment quatre familles multigéniques (*interAE*, *interB*, *interC*, *interD*).

Des amplifications PCR avec des amorces spécifiques des différentes répétitions et de l'ADNr ont permis de valider les jonctions entre blocs EXT et cœurs chromosomiques ainsi

que l'ancrage de ces blocs à l'ADNr. Deux régions subterminales particulièrement tronquées (III β et IX α) ont été reconstituées après séquençage de certains produits de longues PCR.

L'organisation en mosaïque des régions subterminales suggérant de fréquents événements de recombinaison, l'algorithme BLAST a été utilisé pour décrire les sites de recombinaison. La région ADNr et la jonction EXT7/EXT6 montrent des sites « précis » de recombinaison. Des sites « approximatifs » (résolution moyenne de 100 pb) sont régulièrement détectés au niveau des zones de transition entre les blocs EXT et les cœurs chromosomiques. Ces sites approximatifs sont des points chauds de recombinaisons car correspondant à plus de deux types de combinaisons de séquences ADN. Un total de 19 sites de recombinaison a été dénombré. L'absence d'homologie entre extrémités recombinantes témoigne d'une réparation par NHEJ (« non-homologous end joining »). Le caractère mosaïque de la structure des extrémités laisse penser que ce processus de réparation implique un échange d'extrémité chromosomique. Deux arguments nous confortent dans cette hypothèse :

- 1- la conservation de la séquence associée aux télomères dans tous les événements de recombinaison.
- 2- la quasi-absence d'inversion lors des recombinaisons.

La régionalisation chromosomique est clairement marquée par des différences de composition en bases. Le contenu en GC décroît en allant du centre vers chaque région subterminale d'un chromosome donné pour atteindre un minimum de l'ordre de 25%. Cette décroissance est suivie d'une forte augmentation dans l'EXT et l'ADNr (taux moyen : 60%). La modélisation de la dénaturation de l'ADN à 74°C au moyen de l'algorithme Genefizz (interface web à l'institut Pasteur) suggère que les blocs EXT peuvent être assimilées à des régions « fermées » du génome. Ces régions ont un taux de dissociation de la double hélice plus faible que les régions « ouvertes » lorsque l'on modélise un chauffage de la molécule d'ADN. Une analyse plus fine montre une alternance de zones « ouvertes » et « fermées ». Les régions fermées corréleront avec celles contenant les sites de recombinaison.

L'analyse simultanée des biais AT et GC est présentée selon une approche graphique originale. L'ADN est parcouru dans un plan. Un curseur peut, par exemple, se déplacer de gauche à droite selon qu'il rencontre un A ou un T et de haut en bas selon qu'il lit un G ou un C. Les régions centrales des chromosomes de *E. cuniculi* apparaissent très compactes, le déplacement du curseur dans ces régions étant proche du mouvement brownien. Les régions extrêmes, en particulier les blocs EXT, présentent un enrichissement marqué en purines qui pourrait être la signature d'une réplication unidirectionnelle des extrémités chromosomiques.

Ces derniers résultats indiquent que des facteurs tels que la transcription, la réplication et l'hétérochromatinisation influencent la composition en bases de l'ADN des extrémités chromosomiques. Il est remarquable que cette influence concerne, chez *E. cuniculi*, la région de transition entre ADNr et cœur chromosomique, région portant aussi la signature de nombreux événements de recombinaison. Celle-ci méritait donc une appellation spécifique : la « région subterminale ». Il est vraisemblable que les gènes conservés dans un tel contexte chromosomique codent des protéines parasitaires essentielles.

Chapitre 1 : Régionalisation et organisation mosaïque des subtelomères des chromosomes de la microsporidie *Encephalitozoon cuniculi*.

Subtelomere organization in a microsporidian parasite (*Encephalitozoon cuniculi*)

Lavie Laurence^{1,5}, Faye Ngor^{2,5} Dia Ndongo¹, Yeramian Edouard³, Duroure Christophe⁴, Méténier Guy¹, Toguebaye Bhen S.², Vivarès Christian P.¹ and Cornillot Emmanuel^{1,6}

¹ Equipe Parasitologie Moléculaire et Cellulaire, LBP, UMR CNRS 6023, Université Blaise Pascal, Bâtiment Biologie A, 24 Avenue des Landais, 63177 Aubière Cedex, FRANCE

² Laboratoire de Parasitologie Générale, Département de Biologie Animale, Faculté des Sciences et Technologies, Université Cheikh Anta Diop. Dakar, SENEGAL.

³ Unité de Bioinformatique Structurale, Institut Pasteur, 25-28, rue du Dr Roux. 75015 Paris. FRANCE

⁴ Laboratoire de Météorologie Physique, Université Blaise Pascal, 24 Avenue des Landais, 63177 Aubière Cedex, FRANCE

⁵ The two authors contribute equally to this work.

⁶ Corresponding author and present address:

Laboratoire de Biologie Cellulaire et Moléculaire, UFR de Pharmacie, Université Montpellier I, 15 av. Charles Flahault, BP 14491, 34093 Montpellier cedex 5.

Tel: +33 4 67 66 81 25.

Fax: +33 4 67 66 81 29.

E-mail: emmanuel.cornillot@univ-montpl.fr

Abstract

Encephalitozoon cuniculi is a microsporidian parasite harbouring a very small nuclear genome (2.9 Mbp for 11 chromosomes) in which one 16S-23S rDNA unit is associated with each chromosome end. However, the published genome sequence contains only 3 of the 22 expected rDNA units because of the incomplete assembly of subtelomeric contigs. To better characterize the organization of *E. cuniculi* subtelomeres, we undertook the mapping of all duplicated segments upstream of rDNA and the analysis of major biases in nucleotide distribution. Fifteen types of repeated segments (without any gap over one kbp) are detected at the ends of the available chromosome sequences. Some repeat series are shared by several subtelomeres, which let us to distinguish eight homology blocks (EXT1 to EXT8). Long-range PCR and sequencing data validate the junctions with rDNA and allow the re-assembly of missing regions. EXT block and sub-block boundaries reflect 20 recombination sites in which the disruption of sequence homology is suggestive of translocations involving double-strand break repair. Most CDSs of unknown function (122) are ascribed to four potential multigene families. Each transition from chromosome core to subtelomere is marked by an abrupt increase in GC content and the modelling of DNA denaturation shows a relationship between recombination sites and “open-closed” transitions. Two-dimensional representation of GC and TA skews reveals nearly monotonic functions for EXT and rDNA regions, contrasting with multiple loops for chromosome cores. Some similar physicochemical signatures are detected in the subtelomeres of *Guillardia theta*, a eukaryotic endosymbiont also having a reduced genome with one rDNA unit at each chromosome end. The preservation of plastic subtelomeric domains with multigene families in *E. cuniculi* is possibly related to specific roles in parasite adaptability.

INTRODUCTION

Subtelomeres are chromosomal regions that extend upstream of the telomeres and harbour a set of repeated DNA sequences with a variable complexity among eukaryotic organisms (1-3). They are commonly divided into distal and proximal domains. In close proximity of telomeric repeats, the subtelomeric distal domains typically exhibit degenerate telomere-like tracts and some other tandemly repeated sequences. The proximal domains are generally characterized by more or less extensive segmental duplications that may house gene clusters. The dispersal of these homologous blocks among different chromosomes supports the occurrence of ectopic recombination events that should be facilitated by spatial clustering of telomeres at the nuclear periphery (4-5). Some retrotransposon families may be fully integrated to subtelomeres, e.g. Ty5 elements in *Saccharomyces cerevisiae* (6). However, the presence of various multigene families is a more common feature and differential expression of their members is known to play significant roles in the adaptation of microorganisms to environmental changes. This has been particularly documented through studies on antigenic variation processes in human parasites, such as *Plasmodium falciparum* (7), *Trypanosoma brucei* (8) and *Pneumocystis carinii* (9).

The present study deals with subtelomere organization in the nuclear genome of *Encephalitozoon cuniculi*, a species assigned to Microsporidia, a group of fungi-related unicellular eukaryotes which live as obligate intracellular parasites in most animals and form environmentally resistant spores (10). This species can infect a wide range of mammalian hosts and has been recognized as an opportunistic human pathogen in immunodeficient patients (11-12). Consisting of 11 small chromosomes (from 217 to 315 kbp), the *E. cuniculi* genome is only ~ 2.9 Mbp in size (13). As judged by restriction mapping, the extreme regions of all these chromosomes share a common pattern, over approximately 15 kbp, marked by the presence of one 16S-23S rDNA transcription unit (14). Three *E. cuniculi* strains (I, II and III) have been distinguished through some immunological and molecular criteria, especially on the basis of a variable number of GTTT repeats within the unique rDNA internal transcribed spacer separating the two rRNA-coding regions (15). Comparative analyses of isolates differing in host and/or geographic origin have revealed inter- and intra-strain karyotype variability (16-17). Chromosomal length polymorphisms (CLPs) are mostly related to insertion-deletion events (Indels), over 3-10 kbp, having occurred within transition zones between rDNA units and chromosome cores (18). Thus the plasticity of *E. cuniculi* genome appears to be mainly related to DNA rearrangements in subtelomeric areas.

Exemplifying a trend to an extreme genome size reduction among eukaryotic lineages, the genome of *E. cuniculi* GB-M1 (strain I) has been sequenced and found to contain about 2000 potential protein-coding genes that are densely packed and very rarely interrupted by small-sized introns (19). Although single-copy genes are predominant in this miniaturized genome, the two ends of the smallest chromosome (chromosome I) share a common segment of ~ 37 kbp that encompasses the subtelomeric rDNA unit and covers a cluster of six genes (20). This repeated organisation with a dyad symmetry does not seem to be applicable to other *E. cuniculi* chromosomes. However, it should be stressed that the use of “Whole Genome Shotgun” (WGS) approach failed to achieve sequence coverage for most chromosome ends, as indicated by the integration of only three rDNA units into the set of consensus sequences that were submitted to GenBank/EMBL databases: two units for chromosome I and one for chromosome IV (19).

With the working hypothesis that microsporidial subtelomeres may accommodate some multigene families involved in host-parasite relationship, our primary objective was to get a map of all the repeated blocks characteristic for the junctions between rDNA units and chromosomal cores in the *E. cuniculi* genome. After bioinformatics analysis of chromosome consensus sequences, we initiated the reconstruction of lacking regions through a strategy of long-range PCR amplification. Nucleotide distributions (GC content, GC- and AT-skew) along the chromosomes were also examined.

MATERIALS AND METHODS

Purification of *Encephalitozoon* spores and DNA extraction

Encephalitozoon cuniculi GB-M1 isolate was grown *in vitro* in either Madin-Darby canine kidney (MDCK) cells or human foreskin fibroblasts (HFF) at 37°C, as described elsewhere (21). Culture supernatants containing large amounts of parasite spores were pooled then subjected to centrifugation at 3,000 x g for 10 min. The obtained pellet was further treated 15 min at 50°C with 1% SDS to eliminate residual host cell material. Spores were then washed and stored in phosphate-buffered saline (PBS) at 4°C until use.

DNA extraction was performed using the Elu-Quick kit (Whatman Schleicher & Schuell Bioscience, Mantes la Ville, France). About 10⁹ spores were transferred in 120 µl of NaI-containing lysis buffer and incubated for 15 min at 50°C. After mixing with 240 µl of NaClO₄ binding buffer, 60 µl of glass milk solution were added. The suspension was gently

shaken at room temperature for 10 min then centrifuged at 7,000 x g for 30 s. The glass milk was washed in 1 ml of a salt buffer following manufacturer recommendations. Twenty µl of distilled water were added to the recovered DNA. Aliquots of one thousand times diluted solution were used for PCR amplification.

PCR amplification and DNA sequencing

PCR primers are described in Table 1. Long-PCRs were performed using the Triplmaster PCR system (Eppendorf, Le Pecq, France). Amplification started with 10 cycles of a two-step program: DNA melting during 20 s at 94 °C plus annealing/elongation at 68°C with time depending on the estimated amplicon length (3-12 min). Then, 20 cycles of a second two-step program during which the annealing/elongation time is increased by 20 s to each cycle, were applied.

For further partial sequencing, amplified DNA fragments were first partially digested with *Sau3AI* enzyme (New England Biolabs, Ipswich, MA, USA) for 30 min at 37°C then 20 min at 65°C, except for LKIXα3/Lr02r2 amplicon that was restricted with *StyI* because of the rarity of its *Sau3AI* sites. *Sau3AI* fragments were cloned in a *Bam*HI dephosphorylated pUC118 vector whereas *StyI* fragments were cloned in a CIP dephosphorylated empty pGEM-T linearised by *StyI* (New England Biolabs). Sizing was performed between 0.5 and 1.5 kbp. The number of sequenced clones per amplified fragment was 25 for Lr05debrev/Lr02r1-2, 34 for Lr05debrev/Lr13dir3, 47 for Lr08rev/Lr02r1-2, 16 for LKIXα/Lr02r1-2, 63 for LKIXα3/Lr02r2 and 23 for LKIIIβ2/Lr02r1-2. DNA sequencing was performed by the company Genome Express (Meylan, France).

Bioinformatics analysis

Miropeat software (22) was compiled on either UNIX Solaris or LINUX operating system. The source code was distributed by the Washington University server (<http://genome.wustl.edu/pub/software/miropeats>). The known *E. cuniculi* chromosome consensus sequences were compared to each other through the use of three-sequence combinations (Supplementary data, Fig. S1). Blastall program was downloaded from NCBI ftp server or used through the web portal (www.ncbi.nlm.nih.gov/blast). GeneFizz server was used for analyzing base skews along DNA strands (<http://pbga.pasteur.fr/GeneFizz>) (23). The DNA walk algorithm was an automaton reading the sequence from 5' to 3' (24) and was

implemented with a movement going from the (0.0) coordinate toward up, down, left and right for G, C, A and T bases, respectively. Graphic representation was performed using either GNUPlot free software or IDL data analysis and visualization software (<http://www.itvvis.com/IDL/index.asp>).

RESULTS

Most repeated DNA segments are clustered at the ends of *E. cuniculi* chromosome consensus sequences

A search for gap-free repeats in the consensus sequences of *E. cuniculi* chromosomes was achieved using Miropeat software. Only a few small repeats were detected within chromosome cores. By contrast, a significant redundancy of large DNA segments was evident in extreme regions. Fifteen types of repeats (r01 to r05) were discriminated and mapped to the 11 chromosomes as schematized in Figure 1. Repeat numbering was end-to-core organized and sequences were extracted in the same orientation. The two ends of every individual chromosome (Σ) were designated as $\Sigma\alpha$ and $\Sigma\beta$, referring to the 5' and 3' ends of the Watson strand of the sequence deposited in the database, respectively. Precise coordinates of the considered sequences are as found in EMBL annotation files (Supplementary information). The largest repeated segments (2.7-13.6 kbp) mainly exist as 2-4 copies. The extensive duplication on chromosome I (20) is marked by the symmetrical arrangement of five repeats: (I α) r01-r02-r03-r04-r15 / r15-r04-r03-r02-r01 (I β). The r01 repeat (10 kbp) covers the 16S-23S rDNA unit with large flanking regions. Five segments of unique type (be1 to be5) extend between some repeats on three chromosomes, the largest one (be3, 9 kbp) being intercalated between r02 and r09 on VIII β .

Incomplete assembly of subtelomeric regions was evident through the three following observations (i) the rDNA-containing repeat r01 is only re-found on IV α whilst previous physical mapping data have shown that each subtelomere encloses an rDNA unit (14), (ii) no repeats are detected on III β and IX α , and (iii) partial r02 segments at II β , III α , V α , VII α , VIII α , IX β and X β are likely due to artificial truncations. However, it should be noted that a shortened repeat version cannot be considered as an artefact when found in internal position. This is the case of two small r05-type segments bordering an r14 copy on VI β , the three other r14 copies (on II α , VI α and VII α) being flanked by unique DNA sequences.

In a round of sequence comparison, homologous DNA segments were assigned to only eight categories (EXT1 to EXT8) as a result of the clustering of some adjacent gap-free repeats (Figure 1). Because of lacking distal sequences and/or natural rearrangements, the individual blocks within a same category are rather variable in length. For example, EXT4 covers the series r07-r08-r12 (17.7 kbp) on XI α but is apparently restricted to r08-r12 (14.2 kbp) on V β or only r07-r08 (3.4 kbp) on X α . Short unique “be” segments may be associated with one or two adjacent repeats, e.g. r05-be1-r14 for EXT2 and r13-be2-r14 for EXT5. Only the be3 segment on VIII β persisted as a single copy but, waiting for information about missing sequences, this segment was also retained for further analysis and designated as EXT7.

To take into account the fact that every complete *E. cuniculi* subtelomere should present a conserved region encompassing an rDNA unit, the r01 repeat was excluded from EXT1 block. In addition, since the r01-r02 transition was only defined by a small gap on chromosome IV (compared to chromosome I), it was useful to ascertain whether the conserved region may overlap r02. Two breakpoints of sequence homology were detected within the largest r02 version (10 kbp on chromosome I) and positioned at about 3 kb and 5 kb from r01-r02 transition. The breakpoint at 3 kbp was shared by III α , V α and X β , and the second one by VII α and IX β . Two “rDNA-related” subtelomeric segments were therefore distinguished: S1, designating the most conserved segment that largely includes the rDNA unit and extends up to the first breakpoint, with a length of 13.5 kbp (10.5 kbp for r01 + 3 kbp for 5’part of r02) and S2, representing the additional 2-kbp segment bounded by the two breakpoints.

Clearly anchored with the ensemble S1+S2, three EXT1 blocks on I α , I β and IV α were the most representative of complete junctions with chromosome cores. EXT1 block therefore begins with the 3’half of the r02 repeat and can extend on either two or three other repeats (r03-r04 on VIII α , r03-r04-r15 on I α and I β). The r02-r03 boundary has been deduced from a loss of homology between I α and IV α . A feature of the r03 repeat (7.55 kbp) is provided by two CDSs (ECU01_0100 and ECU01_0130 in the case of I α) sharing strong homologies in 5’ coding and 5’ flanking regions (over ~700 bp). The r03-r04 transition is marked by a G insertion on VIII α , at the middle of a 2.9-kbp region devoid of large ORFs and detected CDSs. In contrast, r04 and r15 repeats harbour several duplicated house-keeping genes. An aminopeptidase-dihydrofolate reductase (*dhfr*)-thymidylate synthase (*ts*) gene cluster is indeed specific to r04 repeat. Three other adjacent genes, encoding serine

hydroxymethyltransferase, an ABC transporter and a protein related to the NAP/SET family are characteristic for r15 repeat, restricted to chromosome I.

Long-PCR amplifications provide evidence for linkage of all EXT domains to rDNA units

As a prerequisite to the determination of lacking subterminal sequences among *E. cuniculi* chromosomes, the organization of all EXT blocks and putative anchoring with rDNA were investigated by long-PCR amplification with specific intra-repeat and rDNA oligonucleotides (figure 2). PCR product sizes well fit with predictions from consensus sequences as regard to the junction of rDNA region (S1 + S2) with EXT1 on I α , I β , II β , IV α , VIII α and VIII β ends. We also validated four rDNA-EXT3 junctions (on III α , V α , V β and X β) and two rDNA-EXT5 junctions (on VII α and IX β), as well as EXT-core junctions. For the other chromosomal extremities, the purpose was to determine the connections with EXT2 and EXT4 and the fully missing sequences for III β and IX α . Only one major PCR product was obtained in each experiment based on the use of an EXT2- or EXT4-specific primer. Length conservation of the amplicon suggests a same interval separating rDNA unit from EXT2 on two chromosome ends (II α and IV β) or from EXT4 on six chromosome ends (VI α , VI β , VII β , X α , XI α , XI β). As these amplifications were successful with rDNA primers designed to both sides of the first breakpoint in r02, it is confirmed that the r02 left sequence is omnipresent on *E. cuniculi* chromosome ends. PCR data also indicated that currently known sequences for III β and IX α are located at about 20 and 4 kbp from rDNA oligonucleotides, respectively (Figures 2 & 3).

The four PCR-amplified regions corresponding to unknown junctions for EXT2, EXT4, III β and IX α were partially digested, cloned, and then sequenced. After comparison with genome data, the new sequences were assigned to some EXT sub-blocks (Figure 3) and all the transitions between these sub-blocks were checked by PCR. By this way, we demonstrated that rDNA-EXT2 and rDNA-EXT4 junctions occur via EXT5-1 and EXT3-1 sub-blocks, respectively. The reconstruction of III β revealed an organization quite similar to VI β , through the following sub-block order: (EXT3-1)-(EXT4-1)-(EXT2-2)-(EXT2-3). The organization of IX α , represented by the series (EXT5-1)-(EXT5-2), was related to VII α and IX β , with significant divergence near EXT-core transition. It is noteworthy that the

reassembly of genome reads has confirmed these experimental results. Indeed, no major contigs were really missing, the main problem being to perform unambiguous contig joining.

Twenty hot spots of recombination are found inside EXT blocks

The irregular distribution of subtelomeric repeats among the different chromosomes of *E. cuniculi* (Figure 4) being suggestive of a significant reshuffling of chromosome ends, a third round of sequence comparison (using BLAST algorithm (25)) was performed to detect all the homology breakpoints inside EXT blocks (detailed data are given in supplementary information Figure S2 and EMBL files). Most breakpoints identified in more than one sequence combination, thus appearing as hot spots of recombination. “Precise” sites, for which it was possible to give a point value to the coordinate of DNA exchange, occupy distal positions (closer to rDNA), except for EXT7-EXT6 junction. “Approximate” sites deduced from a significant loss of homology between compared sequences, with a mean resolution of 100 bp, were systematically closer to chromosome cores. Interchromosomal sequence divergence affected the right side of each recombination site and increased when going from rDNA block to chromosome core.

The “precise” sites associated with the rDNA block correspond to the two breakpoints found in the r02 repetition. The first disruption point (R01) is mapped at the middle of a putative CDS, e.g. ECU01_0050 for EXT1 on I α , ECU05_0030 for EXT3 on V α and ECU07_0010 for EXT5 on VII α . The second discontinuity (R02) is representative of the transition from S2 block to either EXT1 or EXT5 and is also located within a putative CDS (ECU01_0070 for EXT1 on I α , and ECU07_0040 for EXT5 on VII α). The other recombination sites found on the different chromosomes carrying EXT1 are of “approximate” type and are described in Supplementary data (Fig. S2A).

We used the same approach to characterize recombination events required for formation of the seven other types of EXT blocks (Figure 5). A total of 20 recombination sites were defined. Some of them may be fused (R10/R19 and R15/R16).

Subtelomeres contain most members of four potential multigene families

Miropeat analysis of different EXT blocks revealed five kinds of repeated tracts, called A, B, C, D and E (Figure 6), most of these repeats overlapping predicted CDSs (Figure S2). In fact, the A and E repeats correspond to the C-terminal and N-terminal regions in a same CDS,

respectively. Moreover, the A and B repeats share sequence similarities and have been assigned at the amino acid level to the same DUF1609 domain in Pfam-A database. The C repetitions are characteristic for putative proteins sharing a DUF1686 domain. This provided the opportunity to classify the repeated CDSs into four potential multigene families called *interAE*, *interB*, *interC* and *interD*.

We counted 19 *interAE*, 32 *interB*, 8 *interC* and 63 *interD* loci, some of them being extrapolated from previous data. While *interAE*, *interB* and *interD* families are restricted to subtelomeres, twelve additional *interC* genes are present in three chromosome cores: 8 for chromosome IX, 2 for chromosome X and 2 for chromosome XI. The transcriptional direction for all *interAE* and *interB* genes is the same as the rDNA unit, i.e. towards the telomere, in contrast with *interC* and *interD* genes in a centripetal orientation. Evidence for transcription capacity has been recently obtained for *interB* genes (26).

The subterminal regions of *E. cuniculi* chromosomes were subsequently analyzed at the nucleotide level to look closely in DNA regionalization.

Local changes in nucleotide composition suggest multiple influences at chromosome ends

The GC% was mathematically calculated using a sliding window (between 100 and 500bp). This approach is useful to study long-distance tendency but is characterized by a low signal-to-noise ratio. As first observed after chromosome I sequencing (20), all the *E. cuniculi* chromosome cores shared a bell-shaped GC% curve with a nearly central apex reaching 54-59% (up to 9 points above genome average). Just above the background noise, it can be noted that the decrease in GC content from the chromosome centre to each sub-terminal region is characterized by oscillations with a 15-20 kbp period. The core sequences making the transition with EXT blocks exhibit the lowest values (25% on average) whereas EXT and rDNA blocks show the highest ones (average at 60%, maximum at 66%). The GC% shift for core-to-EXT transition can be evaluated in terms of maximal variation and slope of this variation ($GC_{max} - GC_{min} / \text{DNA length}$). X α shows the most dramatic shift with a maximal difference of 28.5% and a slope of 3.5 points per 100bp.

To better evaluate the correspondence between genetic data and GC%, Genefizz web interface (<http://pbga.pasteur.fr/GeneFizz/>) gives access to an implementation of the algorithm modelling the dissociation of the two strands of a DNA molecule at a specific temperature, with a window of only 20bp (23). At temperatures between 74 and 78°C, the DNA molecule is incompletely melted, which makes possible discrimination between “open” (dissociated)

and “closed” (double stranded) regions (Figure 7). The higher the GC content, the more the DNA sequence should be closed. This method applied to *Plasmodium falciparum* genome has revealed that closed regions correspond to exons whilst open regions generally correspond to protein-DNA interaction sites such as promoters (27). In the case of *E. cuniculi*, EXT blocks and chromosome cores globally appeared as closed regions. However, more local observations showed that the DNA of each EXT block is clearly marked by an alternation of “open” and “closed” regions. The EXT sites for which it was necessary to increase the temperature of 3°C to obtain a complete melting of the DNA were thoroughly examined. Interestingly, as illustrated in Figure 7, each recombination site coincided with the boundary between an open region and the adjacent closed region in distal position.

A-T and G-C skews along DNA strands were studied simultaneously in a 2D representation, known as “DNA walk” (24). No windows are necessary to calculate the function. Unfortunately, at the present time, the degree of DNA compactness can only be estimated visually. The resolution for the transition between DNA regions differing in A-T or G-C skew is evaluated to ~ 100 bp. The plots obtained for *E. cuniculi* chromosomes support major differences between the textures of core and extreme regions (Figure 8A). Each chromosome core appears as a highly compact area with a wealth of loops, this topology being preserved at various scales. Such a fractal-like structure means that the DNA composition is under the influence of conflicting phenomena. The apparent compactness seems to be higher than in a random sequence measuring Brownian motion. By contrast, EXT and S regions display near-monotonic functions. The transition between the two different topologies fits imperfectly with the analysis of GC content fluctuations. This transition occurs between *dhfr* and *ts* genes within EXT1, in the case of I α , I β and VIII α but has a more internal position (at ~ 20 kbp from EXT region) in the case of II α . The EXT and S blocks diverge as regard to their near-monotonic function: G and A enrichment for EXT block, G and T enrichment for the S block including rDNA.

For comparison, the DNA walk procedure was applied to the chromosomes of some other unicellular eukaryotes including *Plasmodium falciparum*, *Trypanosoma brucei* and several yeasts.. In *Trypanosoma*, the profiles are more reminiscent of *Encephalitozoon* S blocks (figure 8B). The specific GC and AT skew corresponds to a C \rightarrow T transition and has been considered as the result of unidirectional activity of both DNA and RNA polymerases (28). As shown in Figure 8C, *Plasmodium* chromosome ends exhibit some DNA walk properties similar to of *Encephalitozoon* EXT blocks

Subtelomere organization in a eukaryotic endosymbiont (*Guillardia theta*) resembles that in *Encephalitozoon cuniculi*

A comparison with the chromosomes of *Guillardia theta* was especially justified because (i) this organism is an algal endosymbiont with a tiny nuclear genome (555 kbp) reduced to three small chromosomes (180 kbp on average) and (ii) like in *E. cuniculi*, each chromosome end carries one rDNA unit (29). Miropat analysis of *Guillardia* genome detected two contiguous repeats, the largest one (12 kbp) being present on all chromosome ends (Figure 9). These repeats may be viewed as equivalent to *Encephalitozoon* S and EXT regions, with respect to their relative intrachromosomal location and gene composition. Indeed, the segment containing an rDNA unit with the same orientation as in *E. cuniculi* may be assimilated to an S block which is located downstream of a block containing multigene families and thus appearing as an EXT block. Some of these genes are housekeeping genes, e.g. *ubc4* encoding an ubiquitin-conjugating enzyme. All breakpoints in the EXT block are found at 13 ± 1 kbp from the telomere. One sub-block is restricted to three chromosome ends (2α , 2β and 3α) and characterized by the presence of a TFIID encoding gene. Like in *E. cuniculi* genome, sequence diversity increases close to the EXT-to-core transition and this transition is marked by a strong GC% shift. Because of the AT richness of *G. theta* chromosome cores, the shift is sharper than in *E. cuniculi* (Figure 9A). A strong influence on DNA dissociation temperature is evident (Figure 9B) but currently available genetic data are not sufficient to precise some features specifically related to S-EXT junction. However, DNA walk analysis did not show significant difference between the two adjacent subtelomeric regions (Figure 9D). A monotonic curve is associated with each chromosome end and clearly matched with EXT-core transition (G-T enrichment when reading the sequence from core to telomere). Thus, in spite of a higher degree of genome simplification in the endosymbiont, some common signatures are shared by *Guillardia* and *Encephalitozoon* subtelomeres.

Discussion

Genome evolution is commonly characterized by a dynamics of gene duplications and losses. As regard to unicellular eukaryotes, comparative analyses of the genome sequences available for a large number of hemiascomycetous yeast species (*Saccharomyces*, *Kluyveromyces*, *Ashbya*, *Debaryomyces*, *Yarrowia*, *Candida* spp...) have clearly revealed the expansion or contraction of specific gene families in distinct clades (30). Gene copies may originate from various processes including transposon-mediated gene duplications, segmental duplications

and ancient whole-genome duplications. In the case of microsporidia, the knowledge of intra-phylum genome diversity and gene families is very fragmentary. The genome sizes estimated in microsporidian species range between 19.5 Mbp in *Glugea atherinae* (31) and 2.4 Mbp in *Encephalitozoon intestinalis* (32). This indicates that some microsporidian lineages have been subjected to strong genome reduction during evolution. After annotation of the *E. cuniculi* genome sequence, it was evident that the reduction affects the gene content as well as the lengths of intergenic spacers and of genes themselves (19). As here investigated, subtelomeric repeats persist on all chromosome ends of this small genome, which contrasts with the low redundancy of chromosome core-associated genes and lack of large repetitive elements related to transposons. The whole length of the DNA allocated to EXT blocks attains about 300 kbp, thus representing ~ 10% of the haploid genome size. Our long-PCR data have confirmed that, for each chromosome end, these blocks extend upstream of a conserved region centered on one 16S-23S rDNA unit. The mosaic arrangement of individual repeats correlates with the dispersal of the members of four multigene families. Such features are commonly attributed to the proximal subtelomeric domain in other organisms (33). Although mainly available for chromosome I ends, the terminal sequence extending downstream of rDNA is characterized by heterogeneous degenerate tracts of telomeric DNA (20) and can therefore be designated as the distal subtelomeric domain.

Our comparative analyses showed that the dispersal of repeats may be a consequence of numerous recombination events. A greater brittleness of the DNA in the subterminal region, in comparison with other chromosome parts, may justify some physicochemical peculiarities. We indeed showed that this region, although strongly compacted, consists of successive opened and closed areas with some specific DNA sites presenting a very condensed structure. This alternation could represent a favourable context for DNA bending constraints. In such a context, the DNA molecule tends to break more easily. The repair of these breaks would then occur by recombination events between the ends of several chromosomes.

This specific chromosomal region can therefore be considered as the meeting point of unidirectional influences. Both replication and transcription may play a role in a core-to-end direction whereas telosome and heterochromatization act in the opposite direction. Alone, these protein complexes and cellular processes were shown to interact with chromosome extremities and to be either responsible for base composition skew or associated to recombination events. Therefore, we are in favour of a model including physical and spatial

constraints to explain the existence of such a region in chromosomes. Two arguments can be developed.

- 1) The vast majority of recombination events are conservative with respect to the relative orientation of EXT sequence at each chromosome extremities. Only two inversions were indeed observed, the one encompassing a large portion of EXT6 on VIII β and the other being restricted to the CDS ECU07_1800 (Q8SUU3 Uniprot accession) on VII β . This is in agreement with the conception that DNA molecules are oriented in the same way in the nucleus. The orientation of the molecules may result of telomere attachment to nuclear envelope, as yet proposed for *S. cerevisiae* or *P. falciparum* (34-35).
- 2) From telomere to R01 recombination site, chromosome extremities only present strain-specific variations (17). Such level of conservation in subtelomeric regions was also observed in other organisms. The homogenization is the consequence of numerous homologous recombination events. At the opposite, recombination events in the subterminal region are associated to recombination between non-homologous chromosomes. In both *Saccharomyces* and *Plasmodium*, the ends of several chromosomes are thought to form “bouquet” structures during mitosis (34-35). This clustering should then facilitate some unequal exchanges of DNA segments dependent on double-strand break repair processes.

Our work has significantly substantiated subtelomere organization in all the chromosomes of *E. cuniculi* GB-M1 (strain I). This should provide a good basis for further testing the variability of EXT blocks and EXT-associated gene families in isolates representative of strains II and III. Through a first comparative search specifically devoted to the *interB* gene family, we ascertained the presence of *interB* genes in isolates of *E. cuniculi* strain III and of two other *Encephalitozoon* species: *E. hellem* and *E. intestinalis* (26). *InterB* gene sequences were also identified in *Vittaforma corneae* and *Brachiola algerae*, two other human-infecting microsporidia, the latter one having a rather complex genome with a large range of chromosome sizes (up to 2 Mbp). We expected from the *B. algerae* genome project (Genoscope, Evry, France) to have precise answers about the chromosomal context of the conserved *interB* family and the diversity of subtelomeric gene families in a “bulky” microsporidian genome. For a better understanding of the adaptation of microsporidia to their host organisms, it will be especially important to determine whether some of these gene families may be viewed as species- or genus-specific ones and whether their potential expression is required for successful host cell invasion.

Acknowledgements

This work was supported by grants from AUF (Agence Universitaire de la Francophonie; Ngor Faye, invited Professor), EGIDE France (Ndongo Dia PhD) and the Medical Research Foundation (FRM: Fondation pour la Recherche Médicale; Laurence Lavie post-doctoral position).

References

1. Mefford, H.C., and Trask, B.J. (2002) The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet.*, 3:91-102.
2. Britten RJ(1998). Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences. *Proc Natl Acad Sci U S A.*;95(11):5906-12.
3. Rehmeier C, Li W, Kusaba M, Kim YS, Brown D, Staben C, Dean R, Farman M. (2006). Organization of chromosome ends in the rice blast fungus, *Magnaporthe oryzae*. *Nucleic Acids Res.*;34(17):4685-701. Epub 2006 Sep 8.
4. Louis EJ (1995). The chromosome ends of *Saccharomyces cerevisiae*. *Yeast*;11(16):1553-73.
5. Scherf, A., Figueiredo, L.M., and Freitas-Junior, L.H. (2001) Plasmodium telomeres: a pathogen's perspective. *Curr Opin Microbiol.*, 4:409-14.
6. Zou S, Kim JM, Voytas DF (1996). The *Saccharomyces* retrotransposon Ty5 influences the organization of chromosome ends. *Nucleic Acids Res*;24(23):4825-31.
7. Ralph, S.A., and Scherf, A. (2005) The epigenetic control of antigenic variation in *Plasmodium falciparum*. *Curr Opin Microbiol.* 8:434-40.
8. Pays, E., Vanhamme, L., and Perez-Morga, D. (2004) Antigenic variation in *Trypanosoma brucei*: facts, challenges and mysteries. *Curr Opin Microbiol.*, 7:369-74.

9. Stringer JR, Keely SP (2001). Genetics of surface antigen expression in *Pneumocystis carinii*. *Infect Immun.*;69(2):627-39.
10. Metenier, G., and Vivares, C.P. (2001) Molecular characteristics and physiology of microsporidia. *Microbes Infect.*, 3:407-15.
11. De Groote, M. A., G. Visvesvara, M. L. Wilson, N. J. Pieniazek, S. B. Slemenda, A. J. DaSilva, G. J. Leitch, R. T. Bryan, and R. Reves. (1995). PCR and culture confirmation of disseminated *Encephalitozoon cuniculi* in a patient with AIDS: successful therapy with albendazole. *J. Infect. Dis.*;171:1375-1378.
12. Weber, R., P. Deplazes, M. Flepp, A. Mathis, R. Baumann, B. Sauer, H. Kuster, and R. Luthy. (1997). Cerebral microsporidiosis due to *Encephalitozoon cuniculi* in a patient with human immunodeficiency virus infection. *N. Engl. J. Med.* ;336:474-478.
13. Biderre, C., Pages, M., Metenier, G., Canning, E.U., and Vivares, C.P. (1995) Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidium *Encephalitozoon cuniculi*. *Mol Biochem Parasitol.*, 74:229-31.
14. Brugere, J.F., Cornillot, E., Metenier, G., Bensimon, A., and Vivares, C.P. (2000) *Encephalitozoon cuniculi* (Microspora) genome: physical map and evidence for telomere-associated rDNA units on all chromosomes. *Nucleic Acids Res.* 28:2026-33
15. Didier ES, Vossbrinck CR, Baker MD, Rogers LB, Bertucci DC, Shadduck JA (1995). Identification and characterization of three *Encephalitozoon cuniculi* strains. *Parasitology*;111 (Pt 4):411-21.
16. Biderre, C., Mathis, A., Deplazes, P., Weber, R., Metenier, G., and Vivares, C.P. (1999) Molecular karyotype diversity in the microsporidian *Encephalitozoon cuniculi*. *Parasitology*, 118:439-45.
17. Brugere, J.F., Cornillot, E., Bourbon, T., Metenier, G., and Vivares C.P. (2001) Inter-strain variability of insertion/deletion events in the *Encephalitozoon cuniculi* genome: a comparative KARD-PFGE analysis. *J Eukaryot Microbiol.*, Suppl:50S-55S.

18. Brugere, J.F., Cornillot, E., Metenier, G., Vivares, C.P. (2000) Occurrence of subtelomeric rearrangements in the genome of the microsporidian parasite *Encephalitozoon cuniculi*, as revealed by a new fingerprinting procedure based on two-dimensional pulsed field gel electrophoresis. *Electrophoresis*, 21:2576-81.
19. Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretailade, E., Brottier, P., Wincker, P., *et al.* (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*. 414:450-3.
20. Peyret P, Katinka MD, Duprat S, Duffieux F, Barbe V, Barbazanges M, Weissenbach J, Saurin W, Vivares CP (2001). Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora). *Genome Res.*;11(2):198-207.
21. Beauvais B, Sarfati C, Challier S, Dérouin F (1994) *In vitro* model to assess effect of antimicrobial agents on *Encephalitozoon cuniculi*. *Antimicrob Agents Chemother*;38:2440-2448.
22. Parsons, J.D. (1995) Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci.*, 11:615-9.
23. Yeramian E, Jones L (2003). GeneFizz: A web tool to compare genetic (coding/non-coding) and physical (helix/coil) segmentations of DNA sequences. Gene discovery and evolutionary perspectives. *Nucleic Acids Res.*;31(13):3843-9.
24. Lobry JR (1999). Genomic landscapes. *Microbiology Today*; 26, 164-165.
25. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990). Basic local alignment search tool. *J Mol Biol.*;215(3):403-10.
26. Dia N, Lavie L, Metenier G, Toguebaye BS, Vivares CP, Cornillot E (2007). InterB multigenic family, a gene repertoire associated with subterminal chromosome regions of *Encephalitozoon cuniculi* and conserved in several human-infecting microsporidian species.

27. Yeramian E (2000). The physics of DNA and the annotation of the *Plasmodium falciparum* genome. *Gene*;255(2):151-68.
28. Francino MP, Ochman H (1997). Strand asymmetries in DNA evolution. *Trends Genet.*;13(6):240-5.
29. Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG (2001). The highly reduced genome of an enslaved algal nucleus. *Nature.*;410(6832):1091-6.
30. Dujon B (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet.*;22(7):375-87. Epub 2006 May 30.
31. Biderre C, Pages M, Metenier G, David D, Bata J, Prensier G, Vivares CP (1994). On small genomes in eukaryotic organisms: molecular karyotypes of two microsporidian species (Protozoa) parasites of vertebrates. *C R Acad Sci III.*;317(5):399-404.
32. Vivares CP, Metenier G (2001). The microsporidian *Encephalitozoon*. *Bioessays.*;23(2):194-202.
33. Barry, J.D., Ginger, M.L., Burton, P., and McCulloch, R. (2003). Why are parasite contingency genes often associated with telomeres? *Int J Parasitol.*, 33:29-45.
34. Pryde FE, Gorham HC, Louis EJ (1997). Chromosome ends: all the same under their caps. *Curr Opin Genet Dev.*;7(6):822-8.
35. Figueiredo LM, Freitas-Junior LH, Bottius E, Olivo-Marin JC, Scherf A (2002). A central role for *Plasmodium falciparum* subtelomeric regions in spatial positioning and telomere length regulation. *EMBO J.*;21(4):815-24.

Figure Legends

Figure 1. Distribution of the 15 large-sized repeats (r01 to r15) among 20 of the 22 chromosomal extremities in the *Encephalitozoon cuniculi* genome. The r01 repeat has been ascribed to rDNA. The r04 repeat corresponds to dhfr-ts (dihydrofolate reductase - thymidylate synthase) gene cluster. Five DNA segments are of unique type: be1 to be5. Neither r01 nor r02 are present on all extremities. It raised two questions concerning sequence organisation in uncharacterized extremities i) are there more repeats between core and rDNA regions? ii) are there more repeats series? Chromosome numbers are indicated (I to XI).

Figure 2. Experimental validation of consensus end mosaic structure observed after sequence comparison. The chromosome ends were experimentally reconstructed using long-PCR amplification. Amplification products were sometimes larger than 15 kbp. Three hypotheses were tested: EXT-to-core transition (*e.g.* Lr12dir2/LKV β), rDNA anchoring of core extremity (*e.g.* Lr02r1-2/LKIX α) or rDNA anchoring of EXT blocks (*e.g.* EXT4 with Lr01-2/ Lr07rev). The DNA size marker is Lambda cut by *EcoRI* and *HindIII*. Primers sequences are given in Table 1. The right diagram indicates relative position of each primer from R01 recombination site.

Figure 3. Sub-terminal region of chromosome extremity III β . A. PCR strategy. Some of the long-PCR products were sequenced using enzymatic shot-gun. This was useful for determining EXT sub-block composition. The polymorphism was weak except at the EXT-to-core transition (see primers LKIII β 6rev and LKIII β 5 which are specific of the EXT2.2 sub-block at this extremity). The organisation of III β extremity is similar to extremity VI β , but with a longer EXT2 repetition. B. Profiles of PCR products derived from nine primers combinations for the III β extremity structure validation. Oligonucleotides positions are indicated on the top diagram. All the band sizes are in agreement with the predicted structure for this extremity. The marker size used is the 1kb ladder.

Figure 4. Mosaic organization of the 22 sub-terminal regions in *E. cuniculi* genome. rDNA units (S1) are present on all chromosome extremities whereas S2 regions are less frequent. All missing information in Figure 1 (hatched region) has been confirmed by long-PCR experiment and for some extremities by partial sequencing. Detailed description of each extremity is given in supplementary information (Figure S2).

Figure 5. Recombination sites determination. Chromosome extremities can be analysed from extremity to the core junction. The telomeric and distal regions are conserved among all extremities. These regions are associated to the rDNA (*ssu* and *lsu* genes) in the S1 sequence block. The tree organisation of EXT blocks is in agreement with a growing diversity in regions close to the core of the chromosomes. Nodes are recombination sites. Lines represent sub-blocks of EXT sequences. Each line is characterized by the approximate size of the sub-block and the cumulative distance from R01. The R06 recombination site is specific of the end of EXT3-1 sub-block. This was deduced from the analysis of the V α extremity which presents the shortest EXT3 sequence (Figure S2C). But the right sequence of R06 can also be an EXT sub-block: EXT3-2 or EXT4-1. Extremity II α and III β sub-terminal organisations are respectively drawn with bold plain lines (S1-S2-EXT5-EXT3-EXT2) and bold dashed lines (S1-EXT3-EXT4-EXT3-EXT2). The EXT2 block is shared by both extremities (short dashed lines). Respective length information are also surrounded by plain or dashed lines. The loop at the bottom of schema is specific of the EXT6-2 inversion at VIII β extremity.

Figure 6. The EXT blocks contain internal repetitions corresponding to CDS. Miropeat software was used with a threshold of 500. The EXT1 sequence was extracted from extremity I α , EXT2 type sequence comes from extremity II α , EXT3 from X β , EXT4 from XI α , EXT5 from VII α and EXT6 from X α .

Figure 7. GeneFizz analysis of an *E. cuniculi* chromosome end (II α) showing the relationship between open-closed transitions and recombination sites. An overview of complete II α organisation can be obtained by addition of *newECU02* pseudogene sequence (Acc n° AM295151) to VII α consensus. CDSs are represented by boxes with arrows indicating transcription direction and are ordered from telomere (left) to chromosome core (right). Grey colour is applied to the members of four specific multigenic families: *interAE*, *interB*, *interC* or *interD* (Figure 6). Name and Uniprot accession number are given for most CDSs (ECU07_0010 to ECU07_0060 on chromosome VII consensus and ECU02_0010 to ECU02_0100 on chromosome II).

A. Genefizz graphical output using five melting temperatures: 72, 73, 74, 75 and 76°C. Corresponding colours from 72 to 76°C are red, yellow, dark blue, light blue and violet. GC content curve is coloured in green. The EXT region is GC-rich and offers an alternation of

“open” (dissociated) and “closed” (double stranded) areas. The DNA molecule is generally stable up to 71-72°C. An increase of at least 3°C is required for strand dissociation at some loci. Closed segments often match with a region of *interAE* and *interB* genes, encoding a conserved protein domain (PfamA domain of unknown function, DUF1609). GC shift is mapped close to a GAR motif. EXT sub-blocks transitions are marked by recombination hot spots (precise sites: vertical dotted lines; approximate sites: hatched areas). R01 and R02 sites are precisely mapped. R07 site is related to the truncation of an *interB* gene, a rare event for *interB* repository. R19 site corresponds to another rare event as it is fused to R10 (EXT4-EXT2 junction). R08b is a precise recombination site justifying the loss of homology between EXT2 and EXT3 (coord. 5196 on chromosome II). The position of R09 site is less clear because close to Vα EXT-core transition. R20 site is specific to the boundary between EXT2-2 and chromosome core. **B.** Genefizz graphical output using five other melting temperatures: 69, 70, 71, 72 and 73°C.

Figure 8. GC and AT composition skew in chromosomes of *Encephalitozoon cuniculi* and some other eukaryotes.

A. *Encephalitozoon cuniculi* chromosome I. The insert shows whole chromosome analysis. Terminal regions (EXT and rDNA) present biases in base composition. The profile obtained for Iα end is enlarged. The major tendency is represented by an A enrichment from terminus to core. The C enrichment associated with rDNA region extends in the EXT1 block but with AT skew inversion. A series of CDSs is indicated (ECU01_0040 to ECU01_0170).

B. *Trypanosoma brucei* chromosome III (Acc n° CP000066). Genes in kinetoplastid genomes are expressed as large polycistronic transcriptional units. A unidirectional influence of RNA polymerase over several hundreds of kbp may justify a GT enrichment in the non-transcribed strand because of C-> T transition in the non protected strand (28). Twelve regions corresponding to different transcription units correlate with skew inversion, providing a foliated structure. 1, Tb927.3, 100-510; 2, Tb927.3.520-860; 3, Tb927.3.870-1490; 4, Tb927.3.1500-2260; 5, Tb927.3.2270-2490; 6, Tb927.3.2500-3410; 7, Tb927.3.3440-4000; 8, Tb927.3.4010-4890; 9, Tb927.3.4900-5740; 10, Tb927.3.5750-5770; 11, Tb927.3.5780-5830; 12, Tb927.3.5840-5900. A similar phenomenon is observed with bacterial genomes where genes are frequently transcribed in the same direction as the replication of the leading strand (data not shown). Regions 2 and 11 contain ESAG and *vsg* genes.

C. *Plasmodium falciparum* chromosome I (Acc n° AL844501). Chromosome core and end differ in both base composition skew orientation and compactness. The region between

PFA0175w and PFA0205w extending over 30 kbp and containing 10 CDSs is more compact (shorter) than the region between PFA0125c and PFA0170c, 45 kbp long, also containing 10 CDSs. Terminal regions show a more extensive enrichment in purines than internal coding regions, like in the case of *interB* and *interAE* genes within EXT regions of *E. cuniculi*. The slopes of the two curves are identical (the abscise scale chosen for this diagram is different from that in figure 8A).

D. *Yarrowia lipolytica* chromosome A (Acc n° CR382127). The DNA molecule is 10 times longer than *E. cuniculi* chromosome I, but it is fold in a similar space (see abscise and ordinate scales). The 2D representation of AT and GC skew, only reveals a globular structure. Analyses performed on several chromosomes of two other yeasts (*Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*) provide similar results (not shown).

Figure 9. Subterminal regions in *Guillardia theta* chromosomes. A. GC percent analysis. The curve is calculated using a sliding window of 500 bp on the sequence of chromosome 1. The core chromosome sequences are AT rich in *G. theta*. B. Miropeat analysis. One rDNA block containing the telomere, subtelomere and a single copy of the rDNA unit can be found at each extremity. The organisation is very similar to *Encephalitozoon cuniculi*. One EXT block with different recombination sites can also be identified. It contains genes coding either hypothetical proteins or known functions like the ubiquitin conjugating enzyme. These genes belong to multigenic families. C. GC and AT composition skew. Dashed lines indicates the (0,0) coordinate in the diagram. It corresponds to the beginning of the sequence as it is deposited in the databases. The unit per axis corresponds to one curser displacement. Ordinate scale: 200 units. Abscise scales: 200 units for chromosomes 1, 500 units for chromosome 2 and 3 diagrams. The cursor displacement was the same for all the analyses (see chromosome 2).

Table 1. Long-PCR primers.

Figure S1. *Encephalitozoon cuniculi* chromosomes miropeat comparison. A. All chromosome approach. The software options are –inter and threshold at 1000. B. & C. Three-by-three miropeat analyses. The chromosomes were compared and coordinates were extracted from raw data. Present analyses were performed using –inter option and a threshold value of 500. The rDNA-r02-r03-r04-r15 series is established from chromosome I, IV and VIII comparison. The two r15 copies and the farthest similar regions of chromosome I are restricted to this

chromosome. The r02 is entirely present on chromosome I and IV α extremity. It is only partly present on chromosome VIII (both extremities). B. The rpt10-rpt08-rpt07-rpt06 series can be deduced from chromosome VII, X and XI analysis. Two repeats are present on chromosome extremities VII β , X α and XI α (r07+r08). About half of r10 is also present between ends of the XI α consensus and r07+r08 sequence. But the two chromosomes have diverged on this repeat and miropeat identifies only a small signal between chromosome XI and X β extremity. At the opposite repetition r06 is perfectly repeated between X α and XI β extremities. We noticed that like on X α , the r06 copy on extremity VII β remains linked to r07+r08 group of sequences.

Figure S2. Coding DNA sequences (CDSs) and recombination sites (R01 to R18) in EXT blocks distributed among the 22 chromosome ends ($\Sigma\alpha$, $\Sigma\beta$) of *Encephalitozoon cuniculi*. Each CDS is schematized by a large arrow showing the direction of transcription. To gain space, CDS names have been reduced to the last four digit numbers (e.g. 0040 on I α and 0040 on IV α correspond to ECU01_0040 and ECU04_0040 in *E. cuniculi* genome databases, respectively). If available, Uniprot accession numbers for encoded hypothetical proteins are also given (Y103_ENCCU, Y110_ENCCU...). CDSs assigned to four multigene *inter* families (AE, B, C or D) are coloured in grey. Recombination hot spots, indicated by vertical dashed lines (“precise” sites) or hatched zones (“approximate” sites), determine the boundaries between sub-blocks (e.g. EXT1-1, a sub-block of EXT1, is bounded by R02 and R03 sites; see also Figure 5). CDS surrounded by dotted line are in the coding core sequence. CDS with plain lines are EXT are SUB CDSs. Arrows without number are extrapolated from the present data.

A. EXT1 blocks on six chromosome ends (including both ends of chromosomes I and VIII). R01 and R02 recombination sites have been precisely mapped. R03 is an approximate recombination site. Downstream of R01, the sub-block termed S1 harbours one transcription unit for the precursor of large subunit (*lsu*) and small subunit (*ssu*) rRNAs but is only represented in consensus sequences for I α , I β and IV α . EXT1-1 is regularly found and may be adjacent to either EXT1-2 or EXT7. EXT1-2 (from R03 to R04) is partially depicted in the figure. The R03 recombination site is present six times in the genome. It is a hot spot of recombination as the left part of the site (telomere associated sequence S1 and S2 plus EXT1-1) is associated with four different core associated sequences (right part). BlastN alignment between I α and VIII β extremities at this site allowed us to map the EXT1-1-to-EXT7

transition inside one CDS (ECU01-0100 on I α and ECU08_2070 on VIII β). The homology between I α and II β chromosomal extremities is one kbp longer. The loss of homology between IV α extremity and other EXT1 sequences falls between the two previous coordinates, 300 bp away of the EXT1-1-to-EXT7 transition. EXT1-1-to-EXT1-2 transition was thus chosen arbitrarily in an intergenic region. No homologies are found between the different “right” sequences. The extension over about one kbp of the R03 site might be the consequence of either non-homologous recombinations or accumulation of mutations after homologous recombination. EXT1-3 (from R04 to R05) contains genes of known function such as dihydrofolate reductase and thymidylate synthase genes (not shown). For EXT7 organization, see schema **G**.

B. EXT2 blocks on four chromosome ends. The attachment to S2 was validated by PCR experiments and sequencing of Lr05debrev-Lr13dir2 PCR product allowed for missing data reconstruction. EXT2-1 is preceded by EXT5 on II α and IV β . Two newly identified *interB* genes (*newECU02* and *newECU04*) are associated with R07 breakpoint. The VI β EXT4-EXT2 junction could only be approximately determined. In fact, the R10 recombination site is a hot spot of recombination (schema **D**). But this transition imply that a breakage take place at the beginning of EXT2 (R19). It isolates an EXT2 specific sub-block: EXT2-1. The EXT4-EXT2 junction is the fusion of two recombination sites. III β presents the same organisation. Data were obtained after sequencing of PCR product (Figure 3). R08b was precisely determined (coordinate 5196 on chromosome II) on the basis of the comparison with the linkage between EXT3-2 (strongly homologous to EXT2-2) and EXT3-3 (see schema **C**). By contrast, R09 failed to be finely localized because of multiple sequence combinations and degenerated *interAE* gene copies. R20 is specific to the proximal boundary (from EXT to chromosome core). It is noteworthy that EXT-core transitions for II α and IV β are not concomitant of an abrupt increase in GC content.

C. EXT3 blocks on three chromosome ends. EXT3-1 is linked to S1. All individual CDSs of X α diverge significantly from those of III α (note a complete *interC* gene on X α) but a same gene order is preserved. Only the three first CDSs are duplicated on V α . Although EXT3-2 on III α has an organization similar to EXT2-2 (see schema **B**), their whole sequence divergence is sufficiently high to retain an assignment to different blocks. III α and X α sequence are identical in the region of R08a. This site was determined by comparison with EXT2 sequence. R06 is an approximate recombination site associated with an *interD* gene showing a variable

3' region. R13 was mapped to a small part of the region separating the *interAE* gene from the first CDS attributed to chromosome core.

D. EXT4 blocks on seven chromosome ends (including both ends of chromosomes VI and XI). Linkage to S1 occurs via EXT3-1. XI α and V β can be paired and possess the longest form of EXT4 (EXT4-3 is partly shown in the schema). R10 was grossly mapped to the end of an *interD* gene prone to various recombination events. EXT4-2 is a short sub-block that shares strong homology with EXT5-1 and the beginning of EXT1-1 (80 % identity between EXT4-2 on VI α and newly identified EXT5-1 on II α). A unique EXT8 block extends upstream of EXT4-2 on VI α . As for most EXT-core transitions, the placement of R11 was not precise (exchange between *interAE* genes or accumulation of mutations within ECU06_0080 after breakage of intergenic region?).

E. EXT5 blocks on three chromosome ends (including both ends of chromosome IX). The organisation of IX α was extrapolated from long-PCR and sequencing data. EXT5 blocks are distally linked to SUB-2, as for EXT1 and EXT2. EXT5-1 shares 80% identity with EXT1-1, the divergence between these sub-blocks resulting from point mutations and Indels. An abrupt loss of homology was observed between ECU07_0050 (on VII α) and ECU09_2020 (on IX β) that encode the truncated InterB protein Y705_ENCCU (complete gene copies exist on II α and IV β : *newECU02* and *newECU04* genes; see schema **B**). Corresponding R07 breakpoint extends from coordinate 5728 on chromosome VII to a part of the closest intergenic region. Another breakpoint was mapped at the level of the next CDS (ECU07_0060/ECU09_2010) encoding an InterD protein. It was designated as R17 and is EXT5-specific, in contrast with R07 that marks the transition with EXT2-1. EXT5-2 is therefore very short and its origin remains an open question. The EXT5-3-to-core transition (R18) is also very divergent and slightly degenerated on IX α .

F. EXT6 blocks on four chromosome ends. EXT6-1 is just upstream of EXT4-1 on VII β , X α , XI β . Despite that EXT6-2 contains no *inter* genes, its GC content is still high compared to chromosome core sequences. The duplication of EXT6-2 characteristic for X α and VIII α is associated with an inversion (ARF and PRS6B indicate the two terminal genes encoding an ADP ribosylation factor and a proteasome regulatory subunit, respectively). This inversion may be a consequence of a more general genetic event, when considering that (1) on VII β , EXT6-1 displays an inverted *interC* gene (ECU07_1800 coding for Q8SUU3 protein with PfamA domain DUF1686) and (2) on VIII α , EXT6-2 is adjacent to EXT7-2 which appears to

be of chromosome core origin (see schema **G**). R14 was only approximately mapped as it is a hot spot of recombination.

G. EXT7 block on one chromosome end. This unique block is upstream of EXT1-1 on VIII α and is divided into two sub-blocks differing in both gene composition and GC content. EXT7-1 contains a *D-C* gene cluster similar to those in EXT2-2 and EXT3-2 while EXT7-2 harbours seven other CDSs including a gene for ribosomal protein L12 (RL12). GC enrichment is restricted to EXT7-1. The EXT7-EXT6 junction corresponds to the fusion of two recombination sites. R16 carries EXT7 information and its locus is of core origin. The R15 site was first identified as the EXT6-to-core transition (schema **F**). The sequence fusion could be precisely mapped.

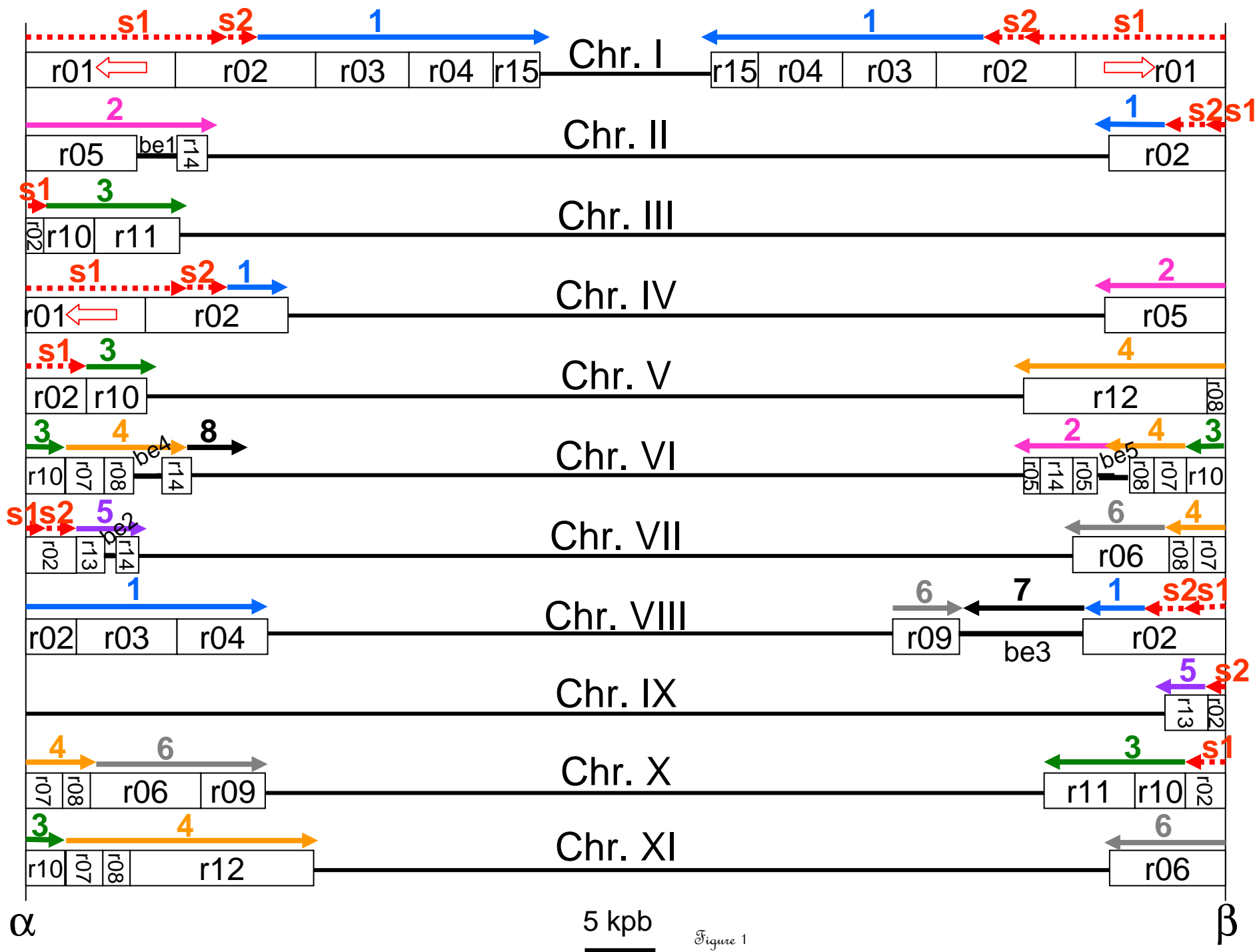


Figure 1

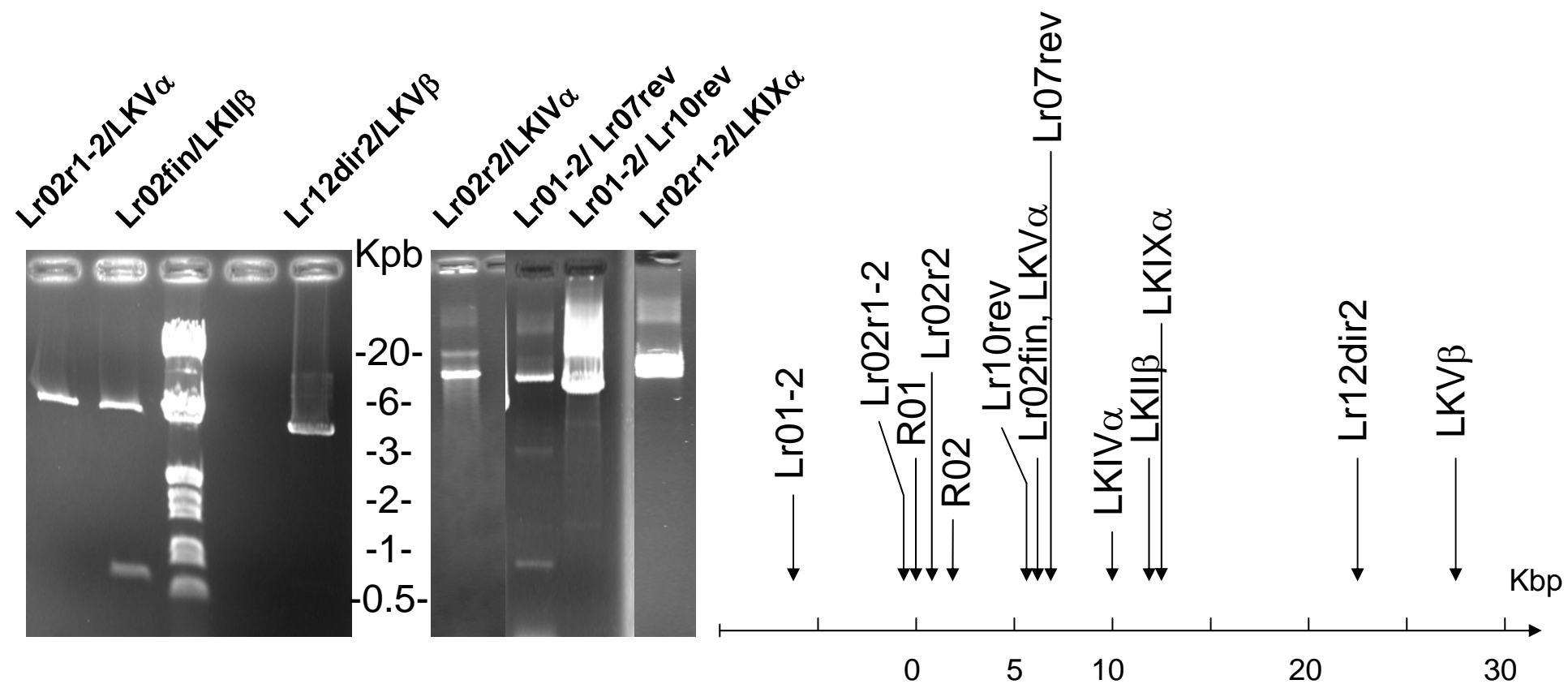


Figure 2

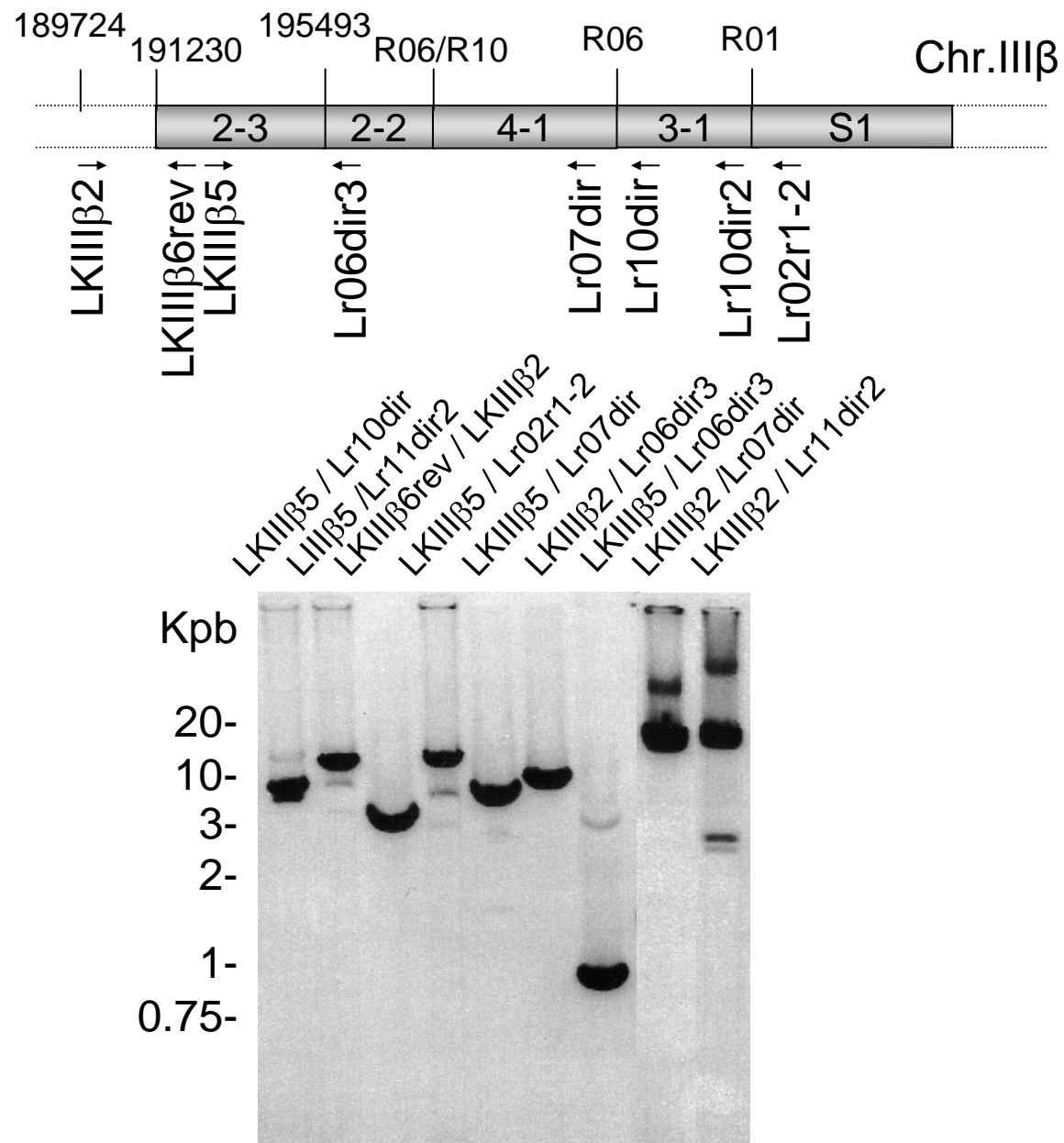


Figure 3

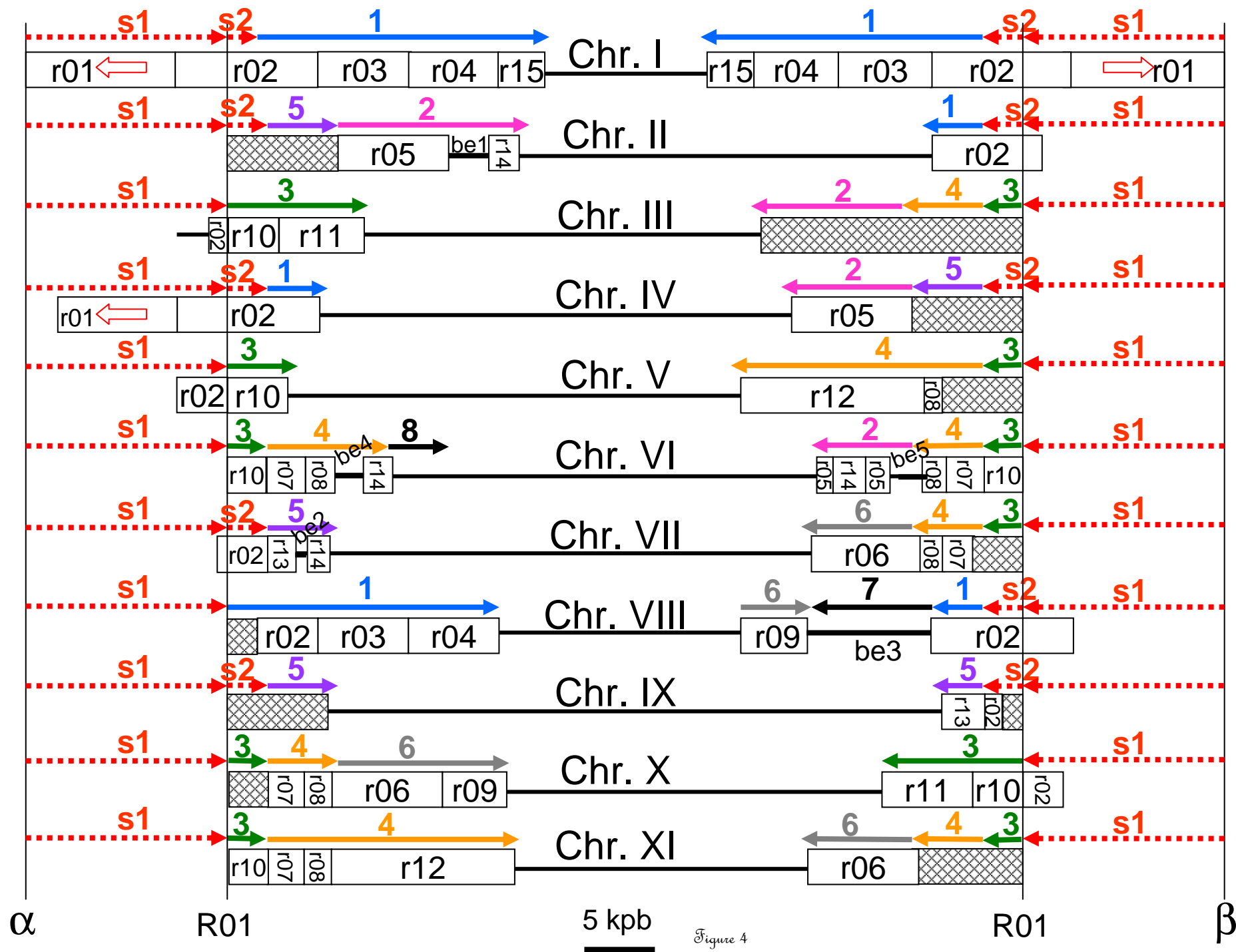


Figure 4

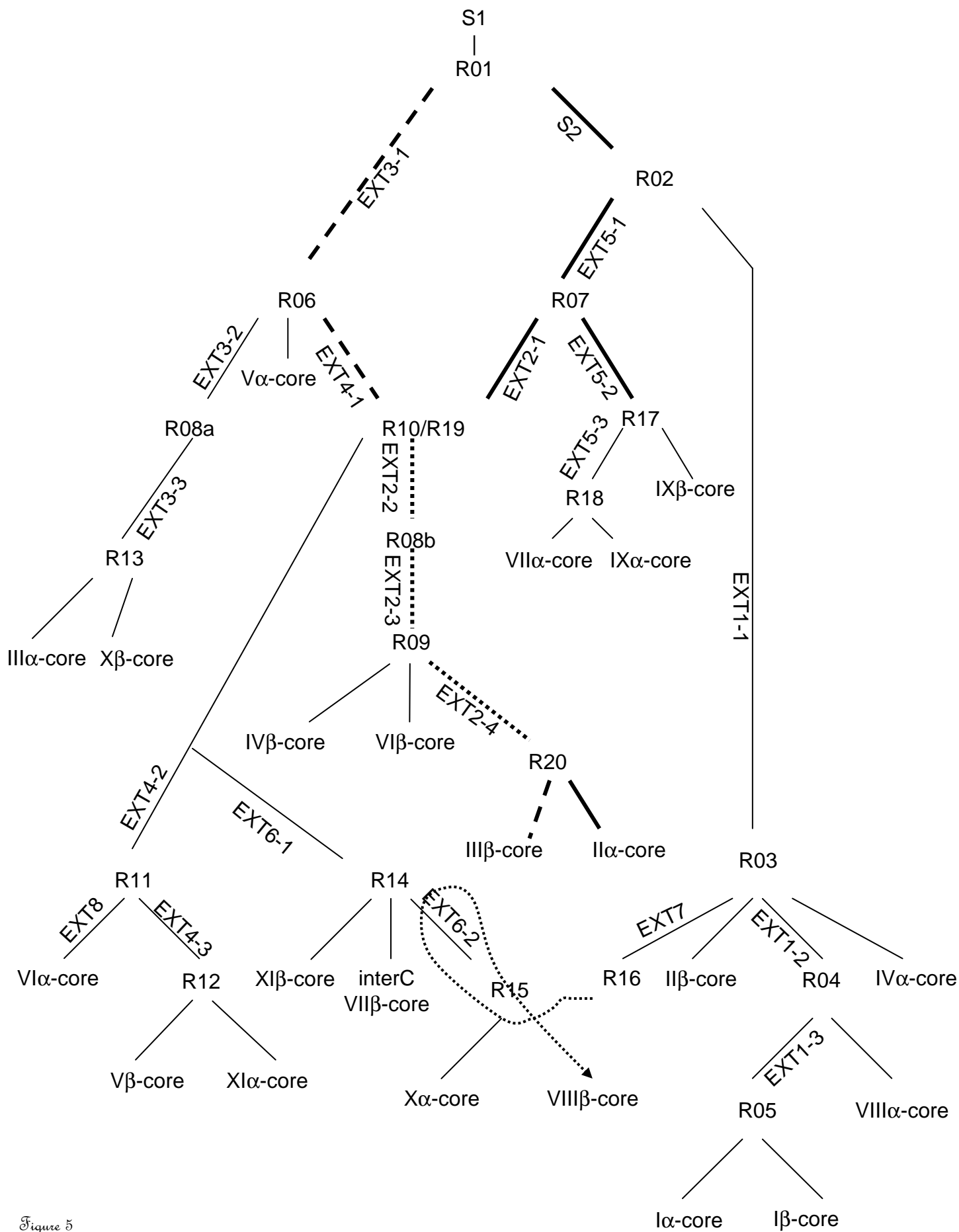


Figure 5

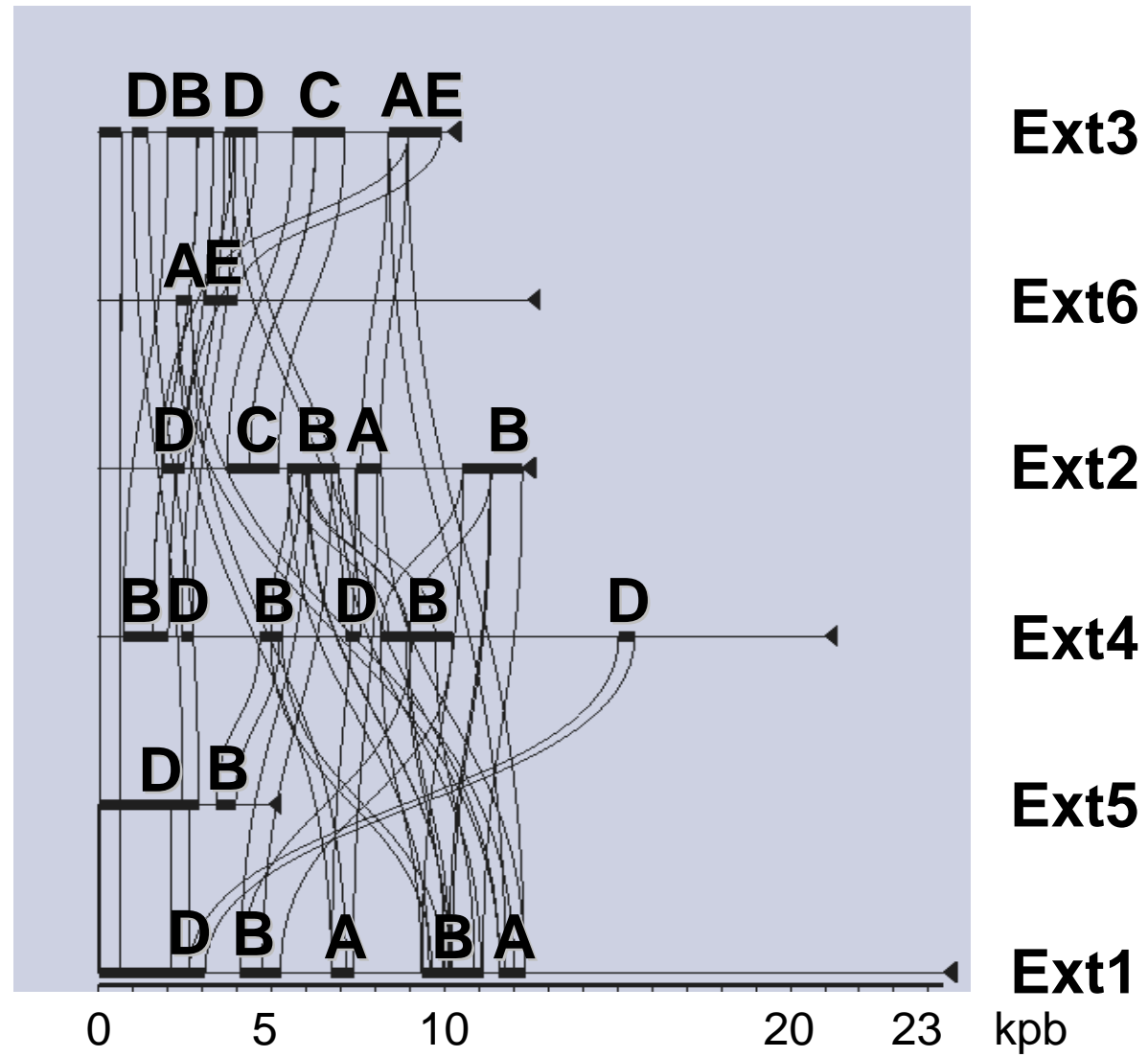


Figure 6

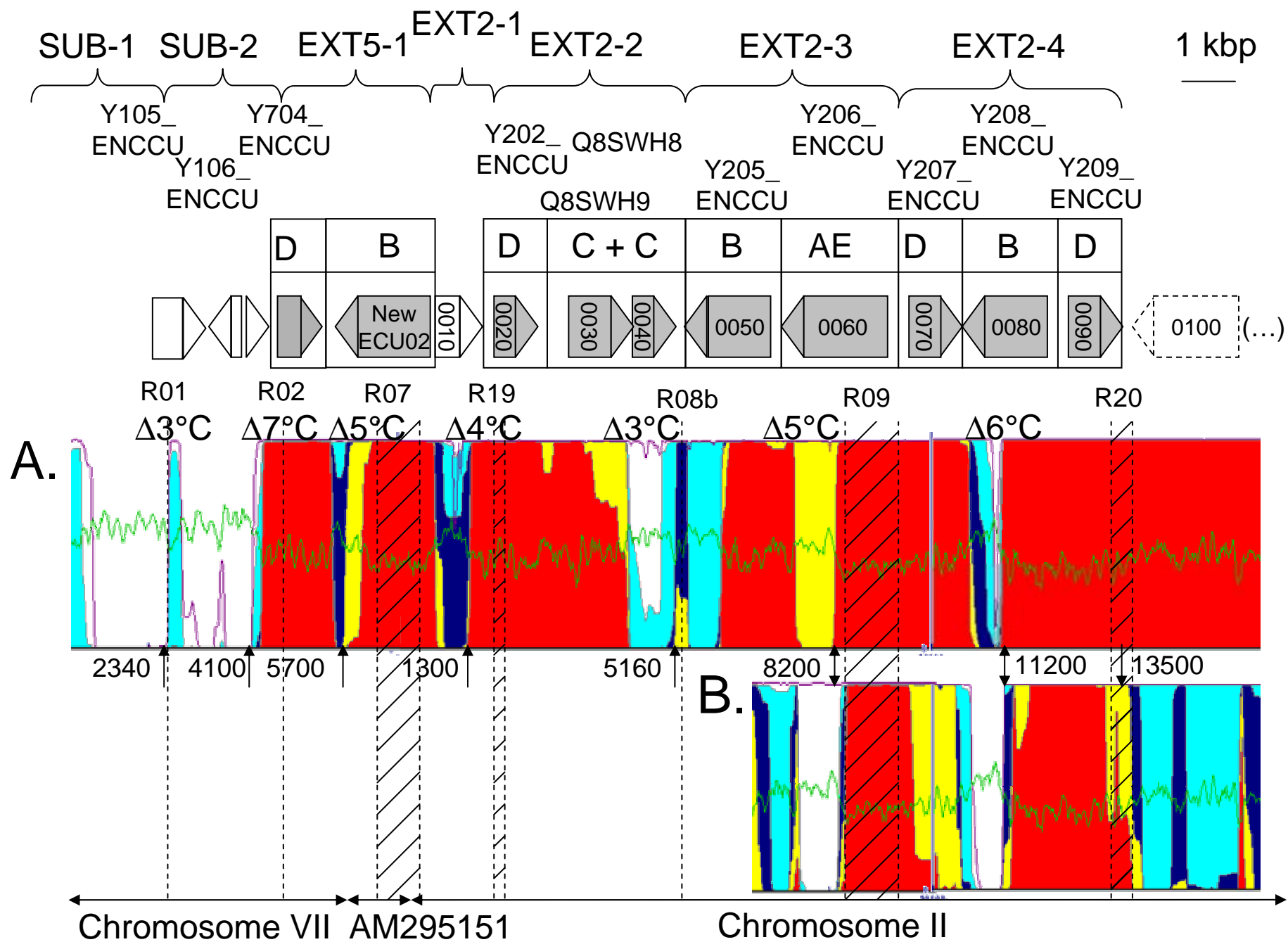


Figure 7

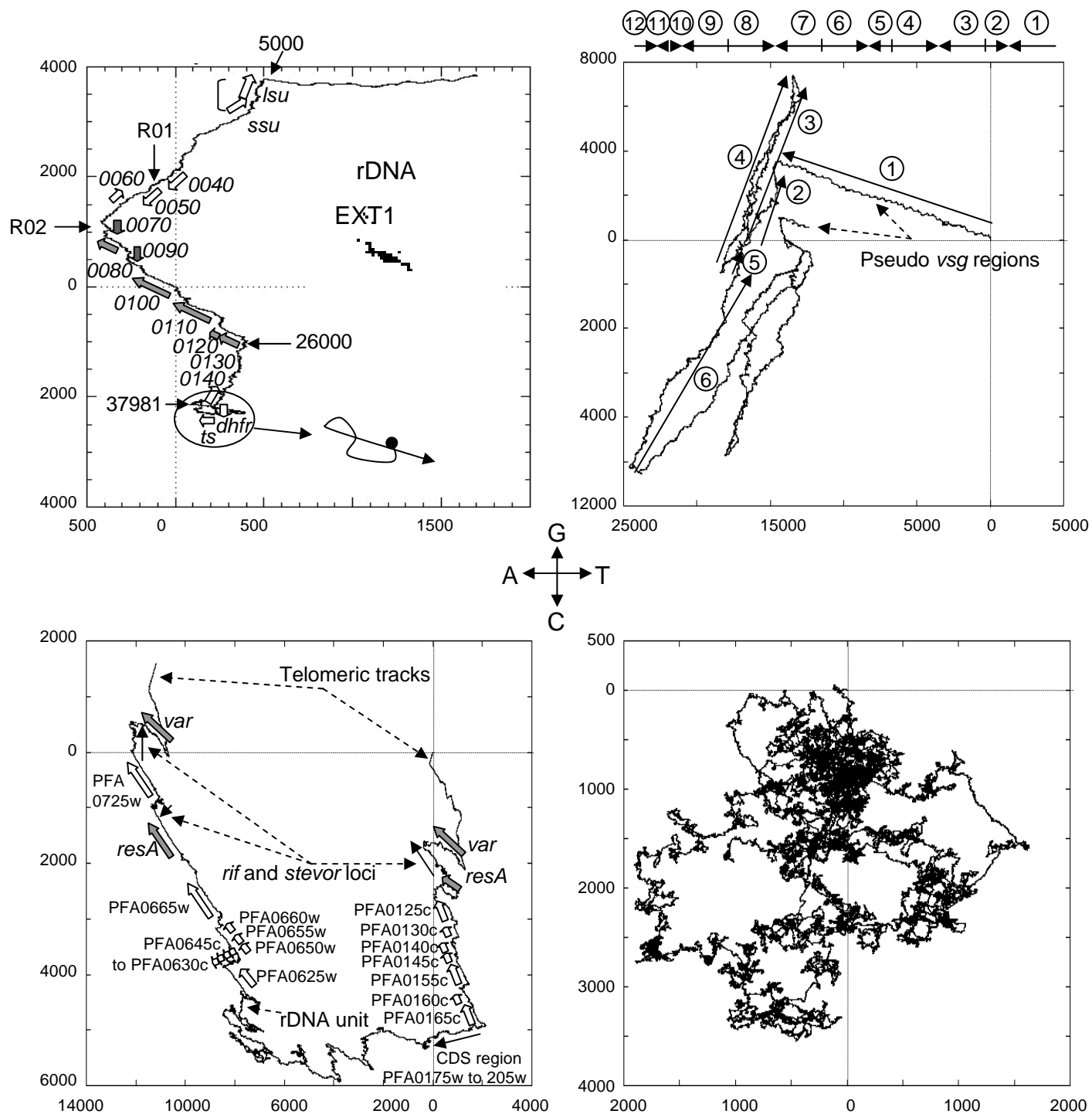


Figure 8

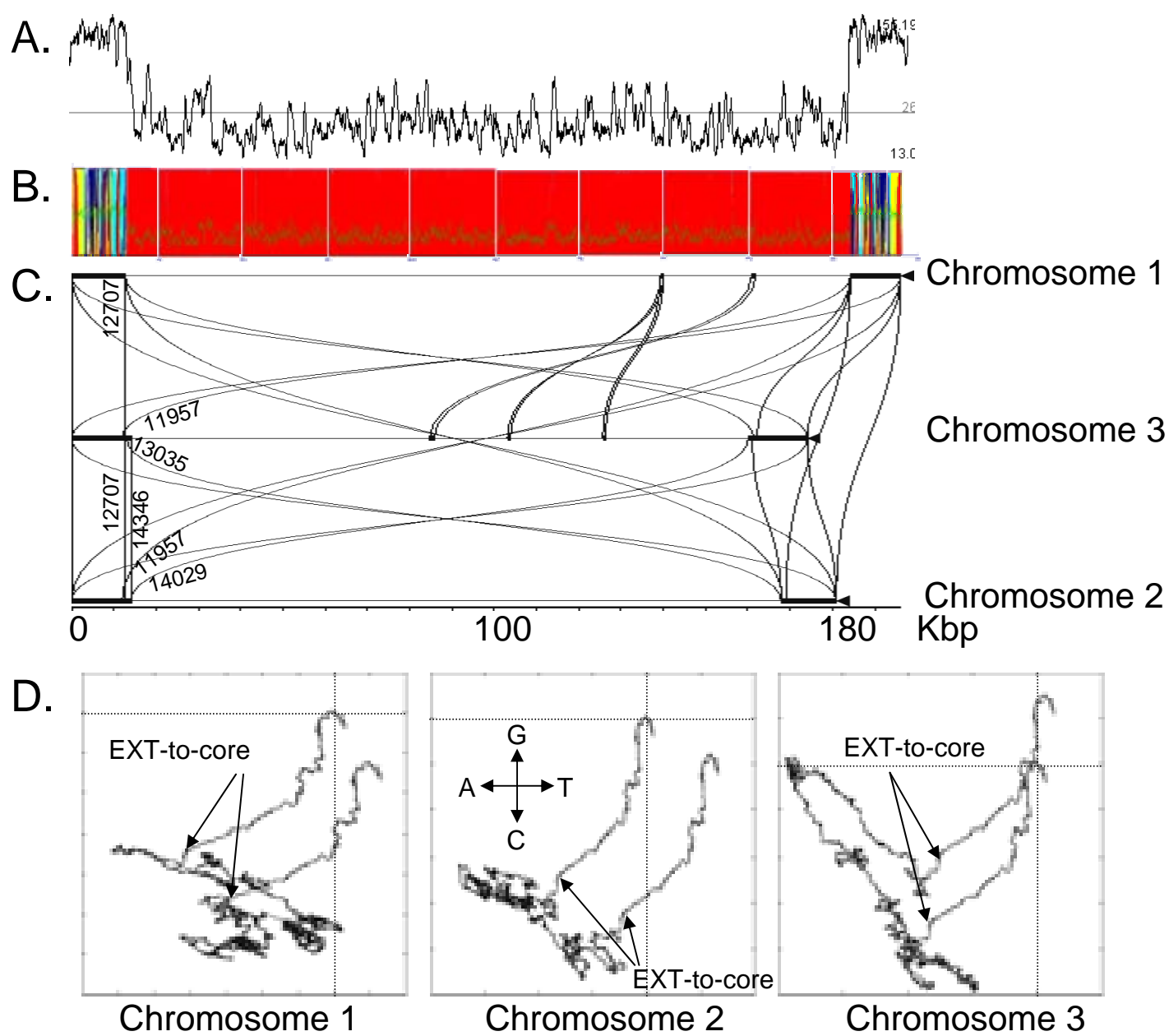


Figure 9

A.

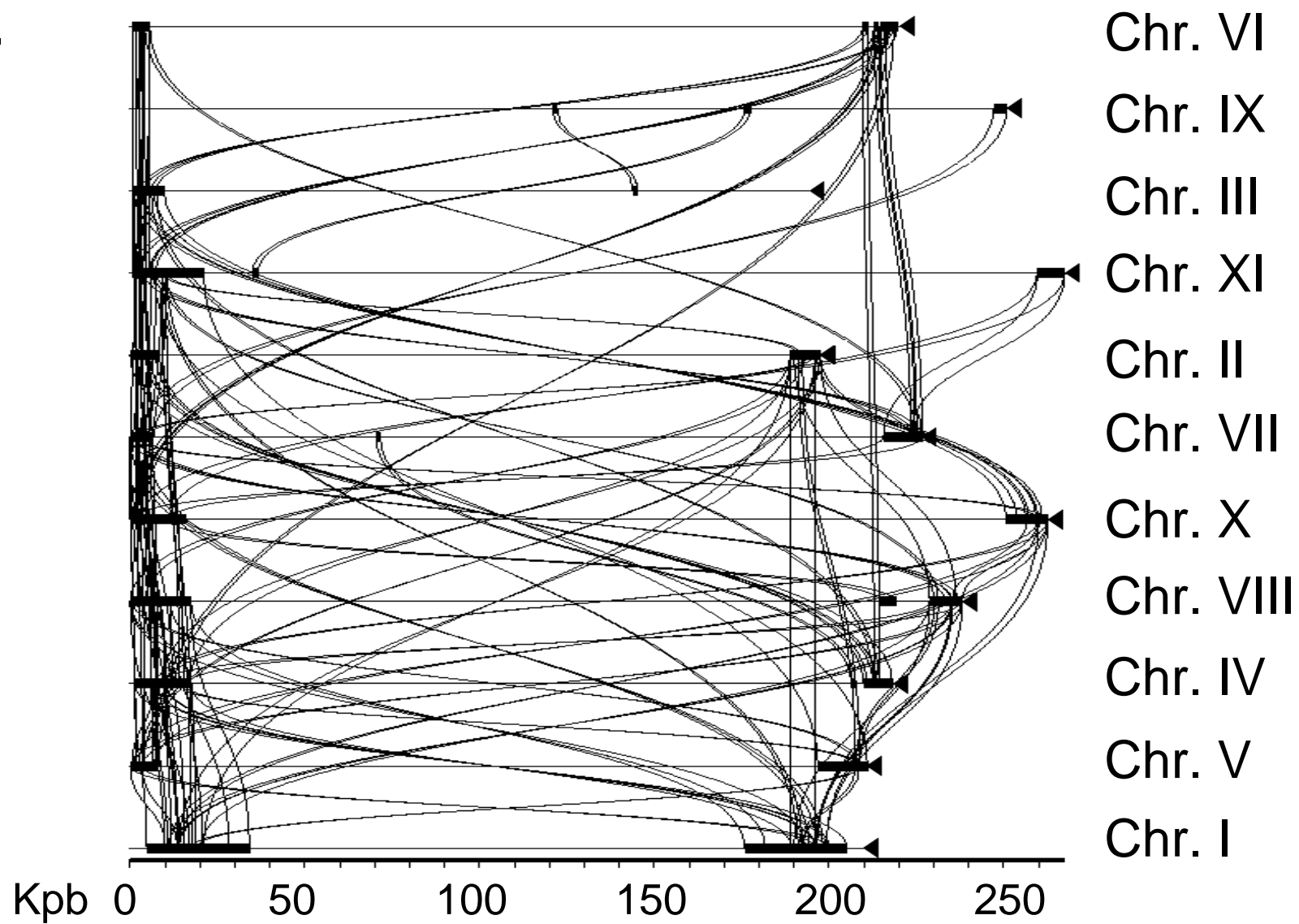


Figure S1

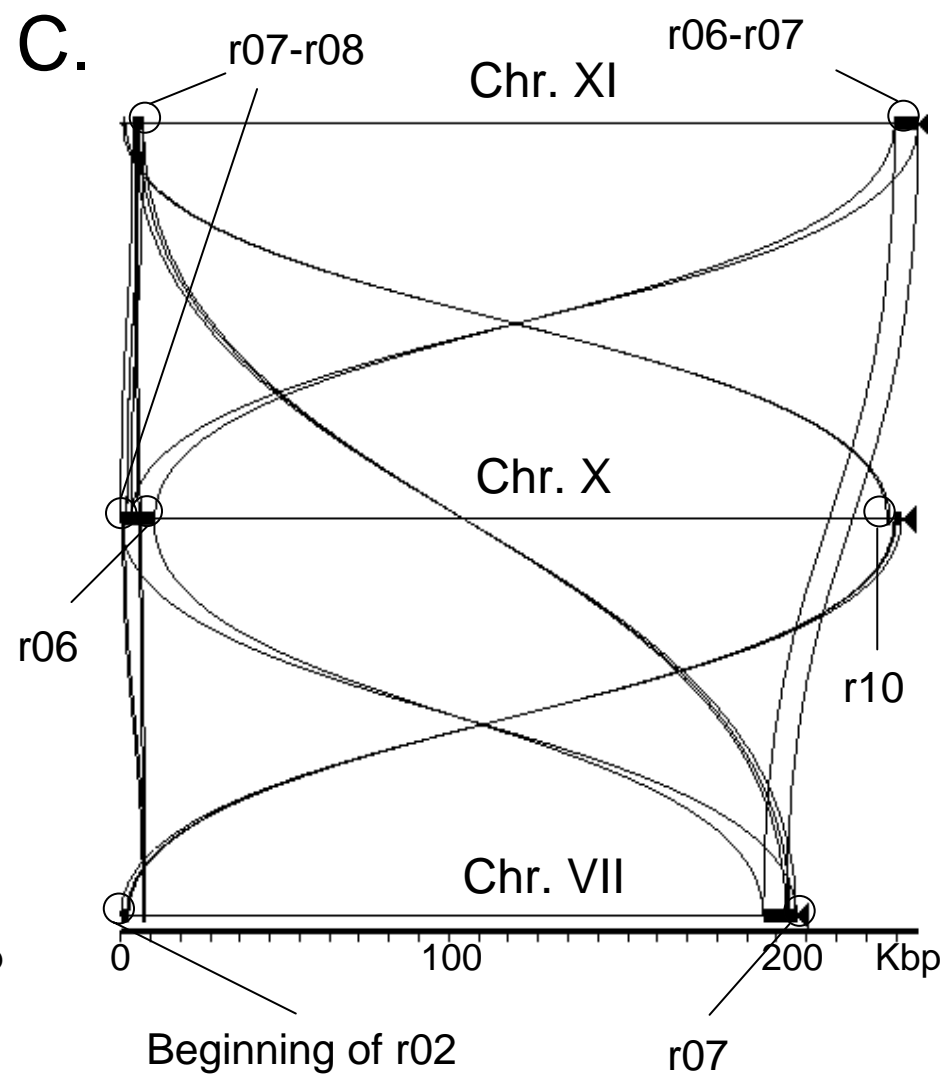
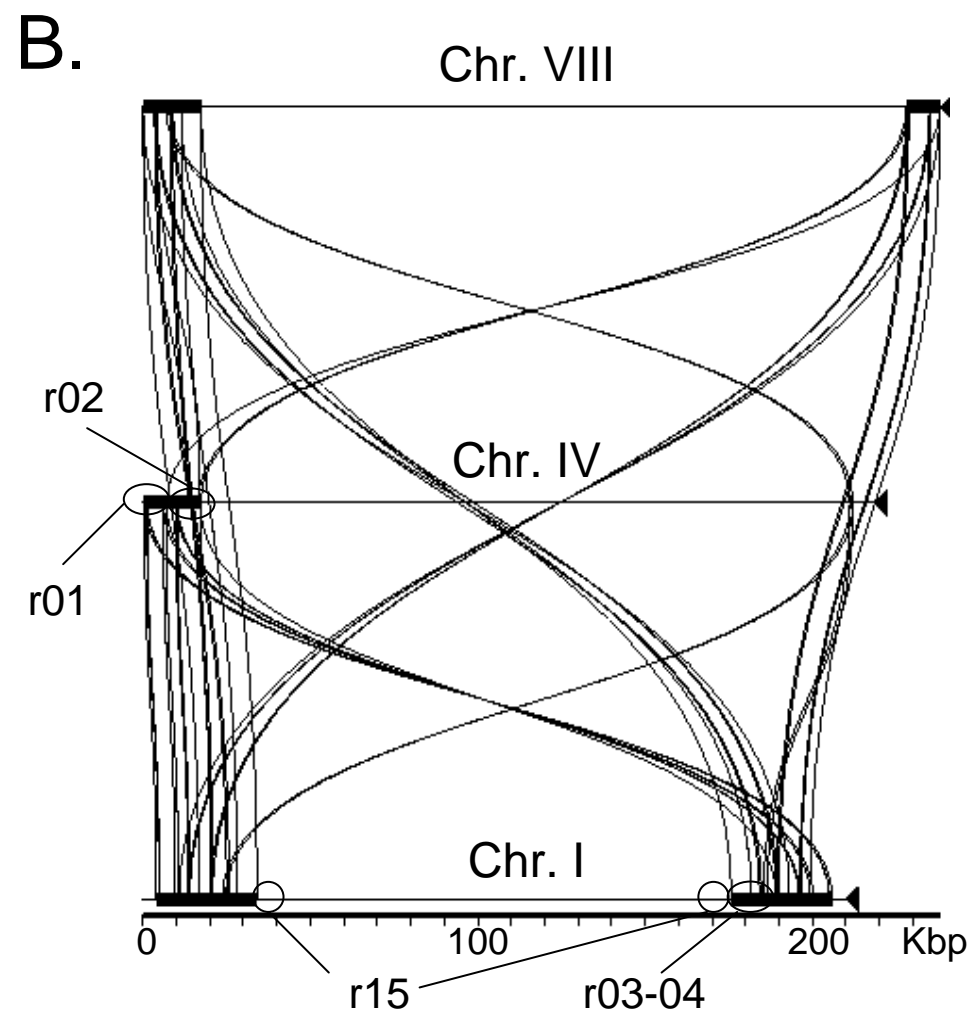


Figure S1

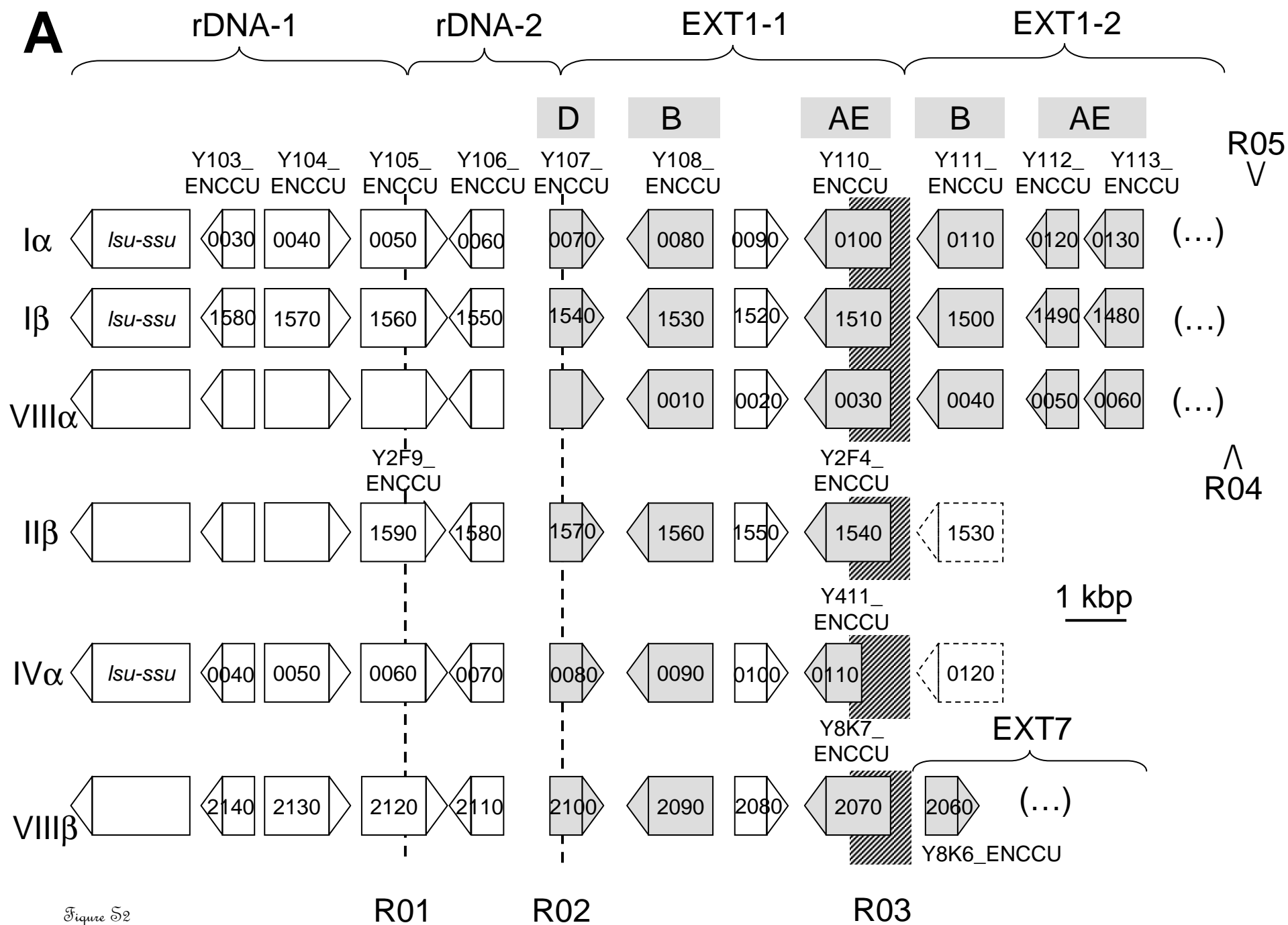


Figure S2

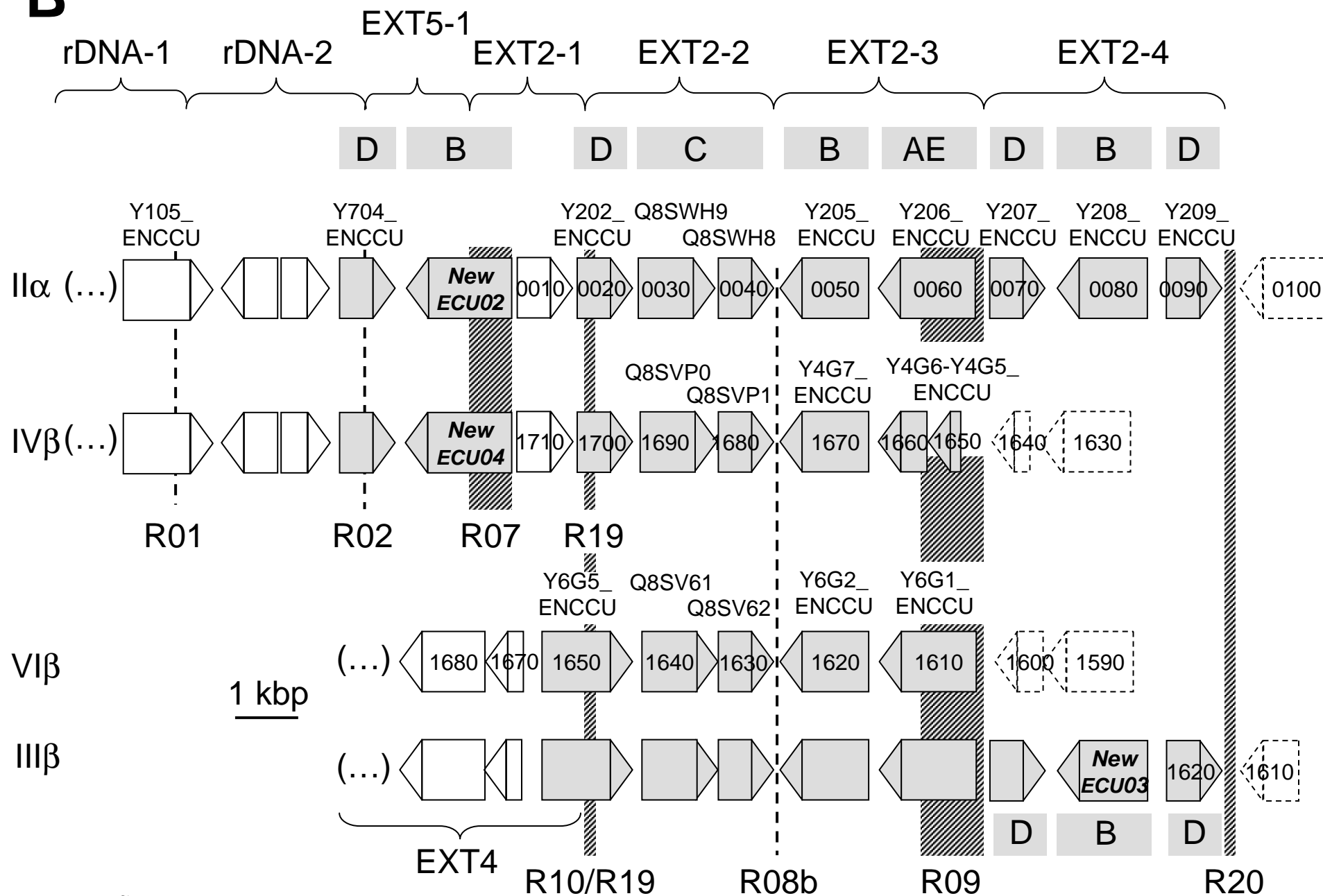
B

Figure S2

C

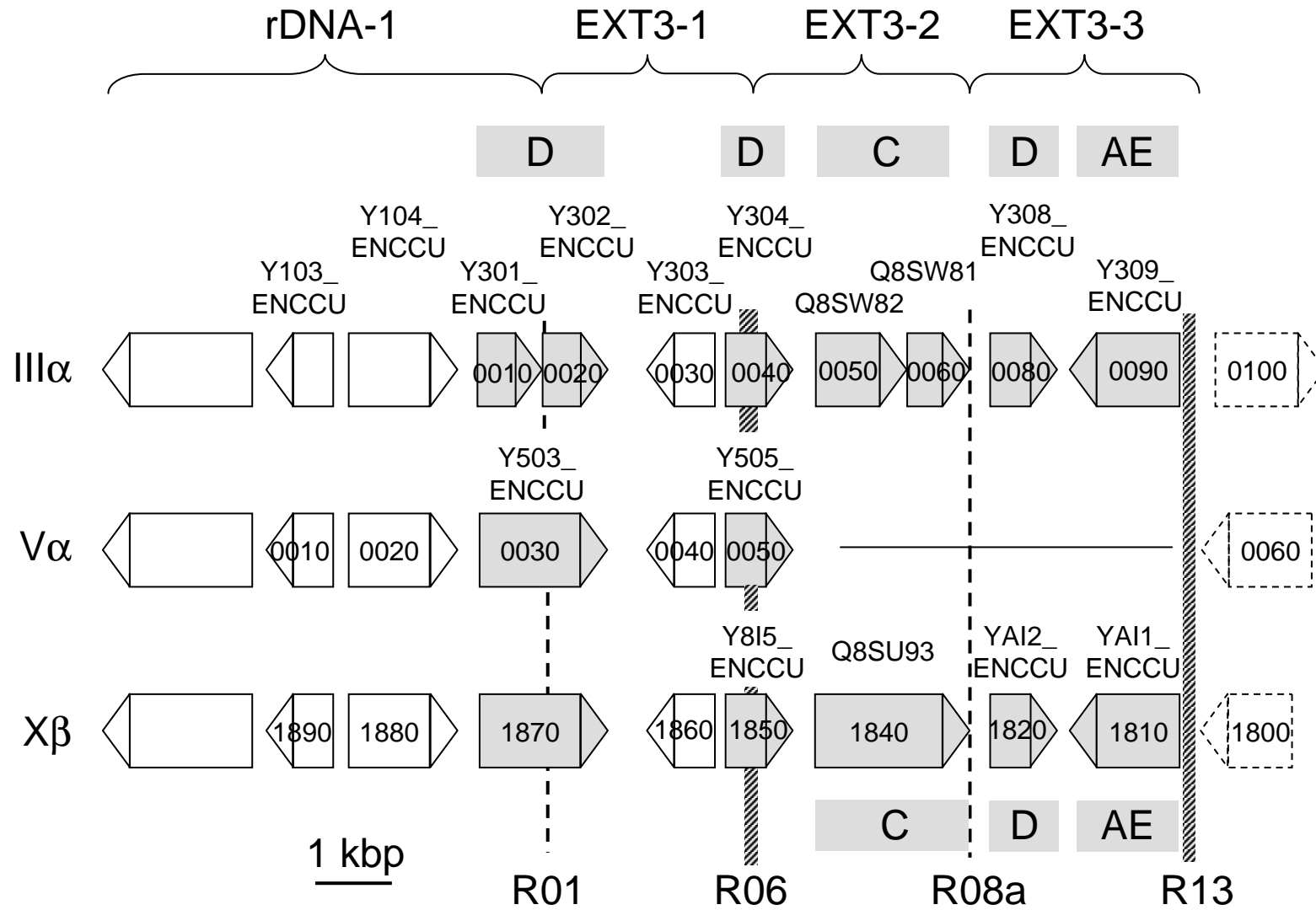


Figure S2

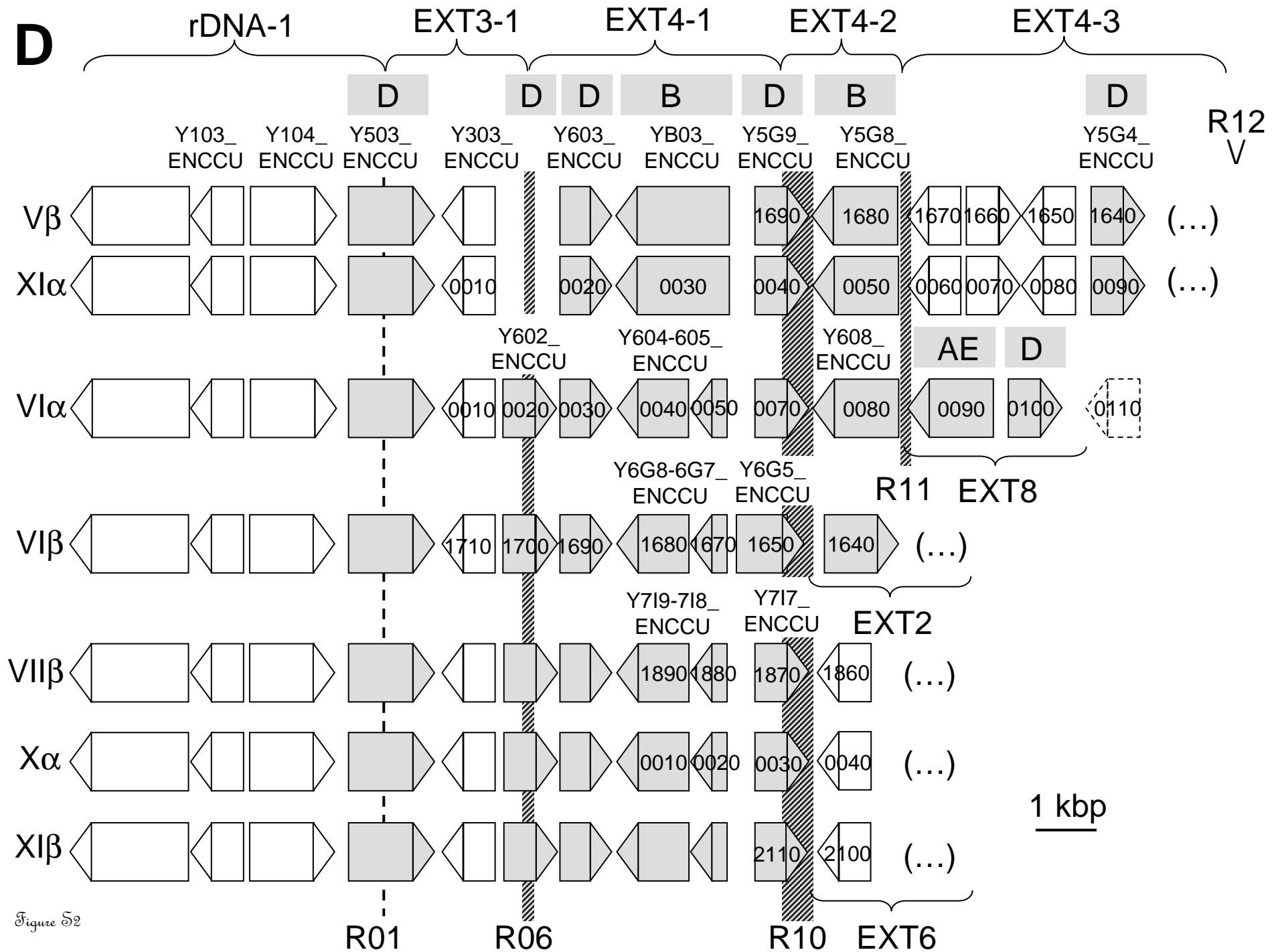


Figure S2

E

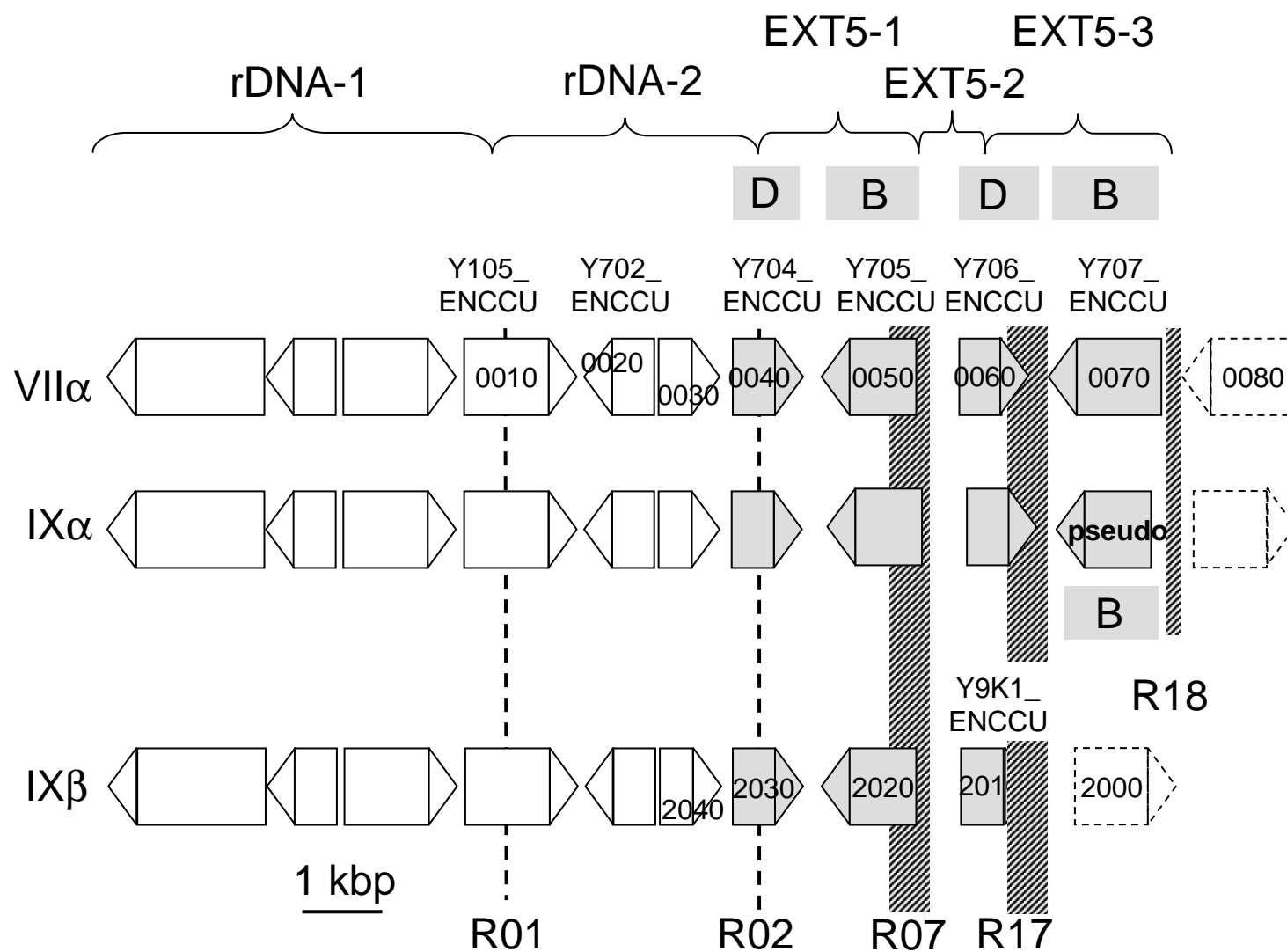


Figure S2

F

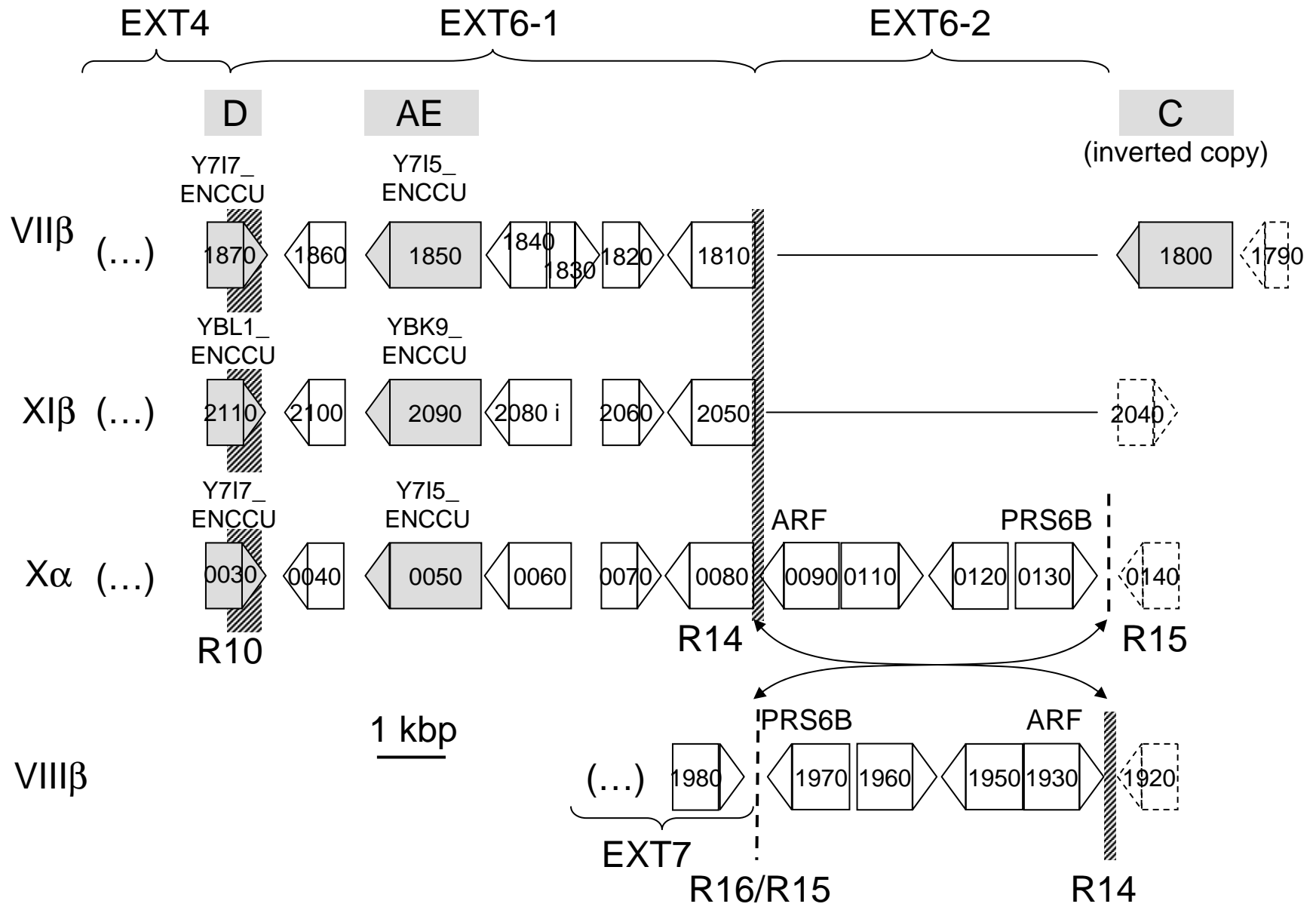


Figure S2

G

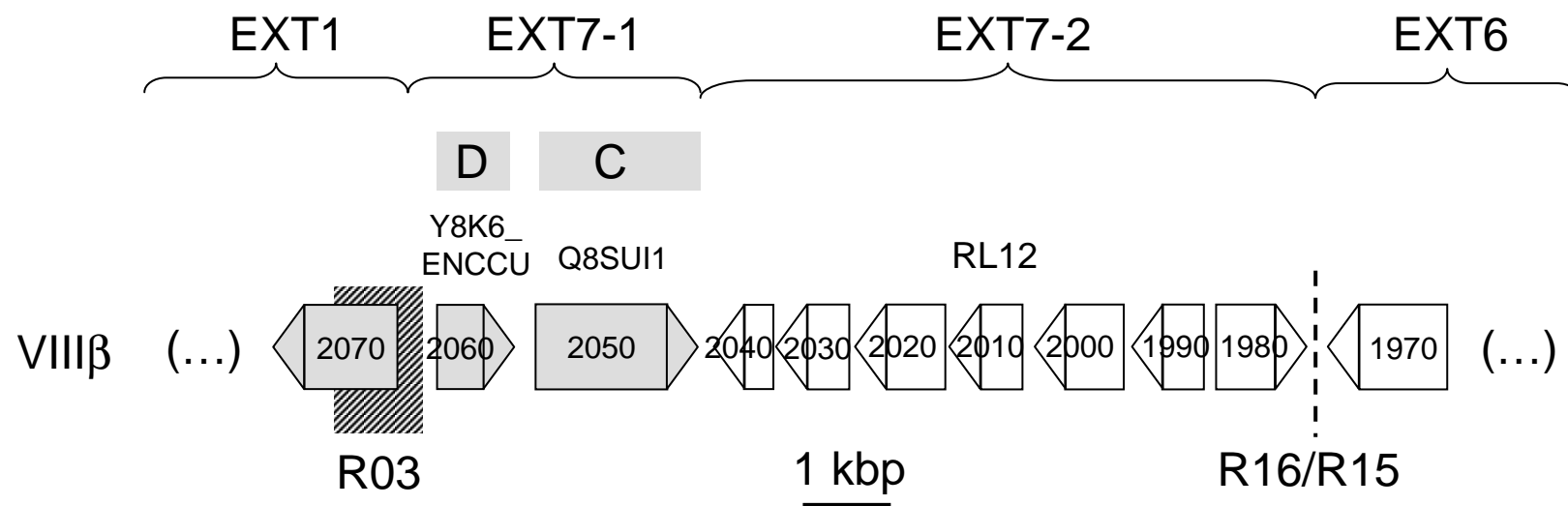


Figure S2

Fichier <ChromosomeI.embl>

```
FT   repeat_region   1..10577
FT                               /label=r01
FT                               /note="Homology with VIa start at position 4035"
FT   repeat_region   1..14126
FT                               /label=S1
FT   repeat_region   10578..20661
FT                               /label=r02
FT   repeat_region   14127..16112
FT                               /label=S2
FT   repeat_region   16113..22172
FT                               /note="homology limits: end of IIb, IVa at 20818 and
FT                               VIIb at 21166"
FT                               /label=EXT1-1
FT   repeat_region   20662..28211
FT                               /label=r03
FT   repeat_region   22173..34125
FT                               /label=EXT1-2
FT   repeat_region   28212..34125
FT                               /label=r04
FT   repeat_region   34126..37941
FT                               /label=r15
FT   repeat_region   34126..37981
FT                               /note="End of homology with Ib: 37941"
FT                               /label=EXT1-3
FT   repeat_region   complement(172039..175857)
FT                               /note="End of homology with Ia: 172042"
FT                               /label=EXT1-3
FT   repeat_region   complement(172042..175857)
FT                               /label=r15
FT   repeat_region   complement(175858..181771)
FT                               /label=r04
FT   repeat_region   complement(175858..187810)
FT                               /label=EXT1-2
FT   repeat_region   complement(181772..189321)
FT                               /label=r03
FT   repeat_region   complement(187811..193870)
FT                               /note="homology limits: end with IIb, IVa at 189165 and
FT                               VIIb at 188817"
FT                               /label=EXT1-1
FT   repeat_region   complement(189322..199405)
FT                               /label=r02
FT   repeat_region   complement(193871..195856)
FT                               /label=S2
FT   repeat_region   complement(195857..209982)
FT                               /label=S1
FT   repeat_region   complement(199406..209982)
FT                               /note="homology with VIa start at position 205948"
FT                               /label=r01
```

Fichier <ChromosomeII.embl>

```
FT   repeat_region   1..2071
FT                       /note="IIa and IVb specific sequence.  Overlapp entry
FT                       AM295151"
FT                       /label=EXT2-1
FT   repeat_region   1..8311
FT                       /label=r05
FT   repeat_region   2072..5196
FT                       /note="Homologous to EXT3-2 from base 3853. 3 prime limit:
FT                       end of homology with VIb (recombination site R08b)."
```

```
FT                       /label=EXT2-2
FT   repeat_region   5197..9390
FT                       /note="End of homology with IVb et VIb: 8259"
FT                       /label=EXT2-3
FT   misc_feature     8312..10523
FT                       /note="unique region between r05 and r14"
FT                       /label=be1
FT   repeat_region   9391..13570
FT                       /note="IIb specific sequence"
FT                       /label=EXT2-4
FT   repeat_region   10524..12251
FT                       /label=r14
FT   repeat_region   complement(187670..193819)
FT                       /note="homologous to chromosome I up to 187777"
FT                       /label=EXT1-1
FT   repeat_region   complement(189271..197426)
FT                       /label=r02
FT   repeat_region   complement(193820..195805)
FT                       /label=S2
FT   repeat_region   complement(195806..197426)
FT                       /label=S1
```

Fichier <ChromosomeIII.embl>

```
FT   repeat_region   1..934
FT                               /label=r02
FT   repeat_region   1..934
FT                               /label=SUB-1
FT   repeat_region   935..4433
FT                               /label=r10
FT   repeat_region   935..4503
FT                               /note="begin homology with VIa and VIB: 2218, XIa:2278"
FT                               /label=EXT3-1
FT   repeat_region   4434..9637
FT                               /label=r11
FT   repeat_region   4504..7453
FT                               /note="End of homology with Va and IVb: 4790. Homologous
FT                               to EXT2-2. Sub-block ended at R08a recombination site (end
FT                               of homology with IIa)"
FT                               /label=EXT3-2
FT   repeat_region   7454..10920
FT                               /note="first loss of homology at position 9579. Specific
FT                               to EXT3."
FT                               /label=EXT3-3
FT   repeat_region   complement(191230..194439)
FT                               /label=EXT2-4
```

Fichier <ChromosomeIV.embl>

```
FT   repeat_region   1..6968
FT                               /label=r01
FT   repeat_region   1..10561
FT                               /label=S1
FT   repeat_region   6969..17096
FT                               /label=r02
FT   repeat_region   10562..12547
FT                               /label=S2
FT   repeat_region   12548..17410
FT                               /note="Homologous to Ia up to coordinate 17256"
FT                               /label=EXT1-1
FT   repeat_region   complement(208800..212929)
FT                               /note="EXT2 specific. End of homology with IIa: 209866"
FT                               /label=EXT2-3
FT   repeat_region   complement(209812..218090)
FT                               /note="region homologous to IIa"
FT                               /label=r05
FT   repeat_region   complement(212930..216051)
FT                               /note="Homologous to EXT3 from base 214270 up to end of
FT                               the sub-block. Similar organisation than EXT3-2."
FT                               /label=EXT2-2
FT   repeat_region   complement(216052..218329)
FT                               /note="EXT2 specific sequence, The extremity is divergent
FT                               with consensus NewECU (entry AM295151) from base 218122"
FT                               /label=EXT2-1
```

Fichier <ChromosomeV.embl>

```
FT   repeat_region   1..495
FT                               /label=r01
FT   repeat_region   1..4063
FT                               /label=SUB-1
FT   repeat_region   496..4064
FT                               /label=r02
FT   repeat_region   4064..8730
FT                               /note="7627: End of homology with IIIa and Xb. 8130:
FT                               EXT3-EXT4 transition"
FT                               /label=EXT3-1
FT   repeat_region   4065..8067
FT                               /label=r10
FT   repeat_region   complement(196790..207110)
FT                               /note="In common with XIa. End of homology: 196833."
FT                               /label=EXT4-3
FT   repeat_region   complement(196830..210398)
FT                               /label=r12
FT   repeat_region   complement(207111..210446)
FT                               /note="End of homology with VIa: 208977"
FT                               /label=EXT4-2
FT   repeat_region   complement(210399..211018)
FT                               /label=r08
FT   repeat_region   complement(210447..211018)
FT                               /note="End with homology to VIIb and Xa (EXT4-EXT6
FT                               transition)"
FT                               /label=EXT4-1
```

Fichier <ChromosomeVI.embl>

```
FT   misc_feature       1..726
FT                               /note="misassembled sequence"
FT   repeat_region      696..3448
FT                               /label=r10
FT   repeat_region      727..3055
FT                               /label=EXT3-1
FT   repeat_region      3056..7497
FT                               /note="End with homology to VIIb and Xa (EXT4-EXT6
FT                               junction)"
FT                               /label=EXT4-1
FT   repeat_region      3449..6135
FT                               /label=r07
FT   repeat_region      6136..7008
FT                               /label=r08
FT   misc_feature       7009..8568
FT                               /note="unique region between r08 and r14"
FT                               /label=be4
FT   repeat_region      7498..10550
FT                               /note="End of homology with XIa and Vb: 8967"
FT                               /label=EXT4-2
FT   repeat_region      8569..9923
FT                               /label=r14
FT   repeat_region      10551..14115
FT                               /note="single copy in E. cucinuli genome"
FT                               /label=EXT8
FT   repeat_region      complement(205505..209768)
FT                               /note="divergent from IIa and IVb type copies"
FT                               /label=EXT2-3
FT   repeat_region      complement(206766..207635)
FT                               /note="correspond to the end of r05 on IIa or IVb"
FT                               /label=r05
FT   repeat_region      complement(207636..209706)
FT                               /label=r14
FT   repeat_region      complement(209707..211184)
FT                               /note="conserved with a central part of r05 type sequence
FT                               on IIa or IVb (closer to IVb)."
```

```
FT                               /label=r05
FT   repeat_region      complement(209770..212819)
FT                               /note="Begin with end of homology with EXT4. Divergent
FT                               copy of EXT2-2. homology with IIa and IVb start in 211109,
FT                               in the middle of CDS ECU06_1640. The R08 recombination is
FT                               difficult to detect. The loss of homology with IIIa and
FT                               Xb is in the interval 212414-211691."
```

```
FT                               /label=EXT2-2
FT   misc_feature       complement(211185..213284)
FT                               /note="unique region between r08 and r05"
FT                               /label=be5
FT   repeat_region      complement(212820..217239)
FT                               /note="End with homology to VIIb and Xa (EXT4-EXT6
FT                               junction). This EXT4-1 extremity is divergent from others
FT                               (divergence starts at position 212975) and EXT4-EXT2
FT                               junction take place 25 bp before EXT4-EXT6 junction"
FT                               /label=EXT4-1
FT   repeat_region      complement(213285..214157)
FT                               /label=r08
FT   repeat_region      complement(214158..216846)
FT                               /label=r07
FT   repeat_region      complement(216847..219599)
FT                               /label=r10
FT   repeat_region      complement(217240..219568)
FT                               /label=EXT3-1
FT   misc_feature       complement(219569..220294)
FT                               /note="misassembled sequence"
```

Fichier <ChromosomeVII.embl>

```
FT   repeat_region   1..2470
FT                               /label=S1
FT   repeat_region   1..3952
FT                               /label=r02
FT   repeat_region   2471..4457
FT                               /label=S2
FT   repeat_region   3953..6742
FT                               /label=r13
FT   repeat_region   4458..6099
FT                               /note="ECU07_0050 encodes a truncated interB gene.
FT                               Complete copy of corresponding interB gene exists on EXT2.
FT                               End of sub-block corresponds to EXT5-EXT2 junction"
FT                               /label=EXT5-1
FT   repeat_region   6100..7278
FT                               /note="end of homology with IXb: 7158. Loss of homology
FT                               with IXb start: 6742"
FT                               /label=EXT5-2
FT   misc_feature     6743..7686
FT                               /note="unique region between r13 and r14"
FT                               /label=be2
FT   repeat_region   7279..9330
FT                               /note="VIIa specific sequence encoding interB gene."
FT                               /label=EXT5-3
FT   repeat_region   7687..8922
FT                               /label=r14
FT   repeat_region   complement(215390..223184)
FT                               /note="End of homolgy with Xa et XIb: 215501."
FT                               /label=EXT6-1
FT   repeat_region   complement(215501..223136)
FT                               /label=r06
FT   repeat_region   complement(223137..224545)
FT                               /label=r08
FT   repeat_region   complement(223185..226576)
FT                               /label=EXT4-1
FT   repeat_region   complement(224546..226576)
FT                               /label=r07
```


Fichier <ChromosomeVIII.embl>

```
FT   repeat_region   1..3875
FT                               /label=r02
FT   repeat_region   1..5386
FT                               /note="ends with homology to IIa. Homology stops in 4032
FT                               for IVa and in 4380 for VIIIB"
FT                               /label=EXT1-1
FT   repeat_region   3876..11425
FT                               /label=r03
FT   repeat_region   5387..17370
FT                               /note="Ia homology stops at 17340"
FT                               /label=EXT1-2
FT   repeat_region   11426..17340
FT                               /label=r04
FT   repeat_region   214721..219069
FT                               /note="map exactly with r09"
FT                               /label=EXT6-2
FT   repeat_region   214721..219069
FT                               /label=r09
FT   repeat_region   complement(219070..223699)
FT                               /note="specific to VIIIB."
FT                               /label=EXT7-2
FT   misc_feature     complement(219070..228052)
FT                               /note="unique region between r02 and r09"
FT                               /label=be3
FT   repeat_region   complement(223700..226559)
FT                               /note="Specific to VIIIB. Present a gene organisation
FT                               similar to EXT2-2 and EXT3-2, but is highly divergent in
FT                               sequence (R08 does not exist)."
```

```
FT                               /label=EXT7-1
FT   repeat_region   complement(226560..232785)
FT                               /note="end of homology with Ia, Ib, VIIIA en 227737, short
FT                               conserved region in 226750"
FT                               /label=EXT1-1
FT   repeat_region   complement(228053..238147)
FT                               /label=r02
FT   repeat_region   complement(232786..234771)
FT                               /label=S2
FT   repeat_region   complement(234772..238147)
FT                               /label=S1
```

Fichier <ChromosomeIX.embl>

```
FT   repeat_region   1..3215
FT                       /note="Encode a pseudo interB gene. Highly divergent from
FT                       VIIa."
FT                       /label=EXT5-3
FT   repeat_region   complement(246100..247826)
FT                       /note="End of homology with VIIa: 248906. Loss of homology
FT                       with VIIa start: 247184."
FT                       /label=EXT5-2
FT   repeat_region   complement(247184..249973)
FT                       /label=r13
FT   repeat_region   complement(247827..249468)
FT                       /note="ECU09_2030 encodes a truncated interB gene.
FT                       Complete copy of corresponding interB gene exists on EXT2.
FT                       End of sub-block corresponds to EXT5-EXT2 junction"
FT                       /label=EXT5-1
FT   repeat_region   complement(249469..251002)
FT                       /label=S2
FT   repeat_region   complement(249974..251002)
FT                       /label=r02
```

Fichier <ChromosomeX.embl>

```
FT   repeat_region   1..2031
FT                               /label=r07
FT   repeat_region   1..3392
FT                               /label=EXT4-1
FT   repeat_region   2032..3440
FT                               /label=r08
FT   repeat_region   3393..11340
FT                               /note="end of homology with VIIb: 11086. end of homology
FT                               with XIb: 11191"
FT                               /label=EXT6-1
FT   repeat_region   3441..11191
FT                               /note="homology with VIIb stops at position 11076"
FT                               /label=r06
FT   repeat_region   11341..15832
FT                               /label=EXT6-2
FT   repeat_region   11484..15832
FT                               /label=r09
FT   repeat_region   complement(249870..253001)
FT                               /note="Specific to EXT3. Loss of homology with IIIa start
FT                               position: 250946."
FT                               /label=EXT3-3
FT   repeat_region   complement(250818..256022)
FT                               /label=r11
FT   repeat_region   complement(253002..255952)
FT                               /label=EXT3-2
FT   repeat_region   complement(255953..259463)
FT                               /note="Sub-block may have been a bit longer. coordinates
FT                               corresponding to end of homology between Va and VIb:
FT                               255667."
FT                               /label=EXT3-1
FT   repeat_region   complement(256023..259463)
FT                               /label=r10
FT   repeat_region   complement(259464..262797)
FT                               /label=r02
FT   repeat_region   complement(259464..262797)
FT                               /label=S1
```

Fichier <ChromosomeXI.embl>

```
FT   misc_feature       1..689
FT                               /note="misassembled sequence"
FT   repeat_region      690..2957
FT                               /note="Beginning in corresponding position 788 on VIa. Sub
FT                               block might be longer as end of homology with EXT3 on Va:
FT                               8411"
FT                               /label=EXT3-1
FT   repeat_region      702..3349
FT                               /note="Begin at corresponding position on VIa: 799"
FT                               /label=r10
FT   repeat_region      2958..7380
FT                               /label=EXT4-1
FT   repeat_region      3350..6018
FT                               /label=r07
FT   repeat_region      6019..7428
FT                               /label=r08
FT   repeat_region      7381..10716
FT                               /note="End of homology with VIa: 8850"
FT                               /label=EXT4-2
FT   repeat_region      7429..20997
FT                               /label=r12
FT   repeat_region      10717..21650
FT                               /note="End of homology with Vb: 20994"
FT                               /label=EXT4-3
FT   repeat_region      complement(259597..267347)
FT                               /note="homology with VIIb stops at position 259712"
FT                               /label=r06
FT   repeat_region      complement(259597..267395)
FT                               /note="End of homology with Xa: 259597"
FT                               /label=EXT6-1
FT   repeat_region      complement(267396..267509)
FT                               /note="end with homology to Va, VIb and XIa."
FT                               /label=EXT4-1
```

Chapitre 2: La famille multigénique *interB*: un répertoire de gènes associé aux subtélomères chez le parasite *Encephalitozoon cuniculi*.

InterB multigenic family, a gene repertoire associated with subterminal chromosome regions of *Encephalitozoon cuniculi* and conserved in several human-infecting microsporidian species

Ndongo Dia, Laurence Lavie, Guy Méténier, Bhen S. Toguebaye, Christian P. Vivarès and Emmanuel Cornillot.

L'analyse de chaque extrémité chromosomique d'*Encephalitozoon cuniculi* a mis en exergue une région d'environ 17 kb incluant toujours une unité ADNr 16S-23S. Dans cette région, nous distinguons successivement le télomère, un domaine subtélomérique très conservé, une unité ADNr puis une « région subterminale » faisant la transition avec le cœur chromosomique et caractérisée par des blocs de séquences répétées (Chapitre 1). Les CDS prédits pour l'ensemble des régions subterminales forment quatre familles multigéniques appelées *interAE*, *interB*, *interC* et *interD*. Leur contexte chromosomique, comparable à celui des familles *vsg*, *var* et *msg* chez d'autres parasites, laisse supposer un rôle dans les interactions hôte-parasite. La famille *interB* fait l'objet de ce deuxième article.

Cette famille comprend une trentaine de CDS pouvant être répartis dans six sous-familles. Le taux d'identité entre les séquences protéiques *InterB* atteint environ 50% dans la partie N-terminale et 70% dans la partie C-terminale. Certaines séquences présentent un peptide signal et des domaines potentiels d'ancrage membranaire. Une région hypervariable (à la fois en séquence et en taille), appelée région « C » car située au centre de la protéine *InterB*, est très riche en glutamine et en lysine. Ces caractéristiques ont permis de développer une approche expérimentale originale pour caractériser qualitativement le répertoire *interB*.

Retrouvé chez *E. hellem* et *E. intestinalis*, il peut être conclu que le répertoire *interB* est conservé au sein du genre *Encephalitozoon*. Un fait remarquable est que ce répertoire est aussi représenté chez deux autres microsporidies parasites potentiels de l'Homme mais de genres différents (*Brachiola algerae* et *Vittaforma corneae*). Par contre, il est absent chez des espèces de 5 autres genres et parasitant strictement des poissons ou des insectes. Bien que le genre *Brachiola* regroupe des parasites de moustiques, des études expérimentales et des cas cliniques ont clairement montré que *B. algerae* peut aussi infester l'Homme et d'autres

mammifères. Si l'on suppose une relation avec une étape de divergence entre *Brachiola* et *Encephalitozoon* remontant à la période où les moustiques ont acquis leur repas sanguin chez les mammifères, les gènes *interB* seraient maintenus depuis au moins environ 200 millions d'années.

Notre analyse du génome de *B. algerae* par des techniques d'électrophorèse à champs alternés à deux dimensions révèle un génome beaucoup plus complexe que celui de *E. cuniculi*, amenant à distinguer trois classes de chromosomes (chromosomes mégabases, chromosomes intermédiaires et mini-chromosomes, comme dans le cas des Kinetoplastidés). La taille du génome haploïde est évaluée entre 15 et 20 Mb. Les résultats d'hybridations révèlent que les gènes *interB* sont localisés sur les extrémités de trois chromosomes intermédiaires et, comme chez *E. cuniculi*, dans une position proche de l'ADNr.

Après extraction des ARN totaux, l'analyse de la transcription par RT-PCR a montré que l'ensemble des sous-familles du répertoire *interB* de *E. cuniculi* est transcrit. Un séquençage aléatoire du répertoire transcrit met en évidence un transcrit majoritaire. Le locus génique correspondant (*newECU03*) est situé sur l'extrémité chromosomique III β et représente l'un des loci *interB* les plus éloignés des télomères. Il est donc tentant d'envisager que l'hétérochomatinisation des régions télomériques et subtélomériques influence le taux de transcription des gènes *interB*. La séquence de *newECU03* étant très similaire à celles des gènes *interB* du chromosome II, sa transcription préférentielle pourrait être la conséquence d'un évènement de recombinaison associant un locus *interB* à un site d'expression. Une dépendance vis-à-vis du stade de développement parasitaire est une autre hypothèse à éprouver.

La conservation de gènes *interB* chez plusieurs microsporidies parasitant l'Homme permet d'envisager des applications au diagnostic et à l'épidémiologie des microsporidioses, pouvant s'appuyer sur les stratégies PCR ici mises en oeuvre.

***InterB* multigenic family, a gene repertoire associated with subterminal chromosome regions of *Encephalitozoon cuniculi* and conserved in several human-infecting microsporidian species**

Ndongo Dia · Laurence Lavie · Guy Méténier ·
Bhen S. Toguebaye · Christian P. Vivarès ·
Emmanuel Cornillot

Received: 15 September 2006 / Revised: 8 November 2006 / Accepted: 28 November 2006
© Springer-Verlag 2007

Abstract Microsporidia are fungi-related obligate intracellular parasites that infect numerous animals, including man. *Encephalitozoon cuniculi* harbours a very small genome (2.9 Mbp) with about 2,000 coding sequences (CDSs). Most repeated CDSs are of unknown function and are distributed in subterminal regions that mark the transitions between subtelomeric rDNA units and chromosome cores. A potential multigenic family (*interB*) encoding proteins within a size range of 579–641 aa was investigated by PCR and RT-PCR. Thirty members were finally assigned to the *E. cuniculi interB* family and a predominant *interB* transcript was found to originate from a newly identified gene on chromosome III. Microsporidian species from eight different genera infecting insects, fishes or mammals, were tested for a possible intra-phylum conservation of *interB* genes. Only representatives of the *Encephalitozoon*, *Vittaforma* and *Brachiola* genera, differing in

host range but all able to invade humans, were positive. Molecular karyotyping of *Brachiola algerae* showed a complex set of chromosome bands, providing a haploid genome size estimate of 15–20 Mbp. In spite of this large difference in genome complexity, *B. algerae* and *E. cuniculi* shared some similar *interB* gene copies and a common location of *interB* genes in near-rDNA subterminal regions.

Keywords *Encephalitozoon cuniculi* · *Brachiola algerae* · Chromosome extremity · Multigenic family · KARD 2D-PFGE

Introduction

Microsporidia are a fungi-related group of single-celled, obligate intracellular parasites that form spores and invade a wide variety of animals. Several genera, including *Encephalitozoon*, have emerged as agents of opportunistic infections in AIDS patients (Desportes et al. 1985). Some infections of microsporidian origin have also been reported in the immunocompetent persons (Didier et al. 2004). The genome of *Encephalitozoon cuniculi*, only 2.9 Mbp in size, was the first to be sequenced among eukaryotic parasites (Katinka et al. 2001). This genome consists of 11 linear chromosomes, each chromosome extremity being marked by the presence of a 16S-23S rDNA copy (Brugere et al. 2000a). However, chromosome extremities failed to be fully assembled in the final release of the genome sequence. Telomeres and subtelomere repeats were poorly defined and only three of the 22 rDNA copies were assembled. A comparative analysis of the available *E. cuniculi* extreme sequences has revealed a specific

Communicated by S. Michaeli.

The members of the genus *Brachiola* have been transferred to the genus *Anncaliia* [Franzen et al. (2006) J Eukaryot Microbiol 53:26–35].

N. Dia · L. Lavie · G. Méténier · C. P. Vivarès ·
E. Cornillot (✉)
Equipe Parasitologie Moléculaire et Cellulaire,
Laboratoire de Biologie des Protistes, UMR CNRS 6023,
Bâtiment Biologie A, Université Blaise Pascal,
63177 Aubière cedex, France
e-mail: Emmanuel.CORNILLOT@univ-bpclermont.fr

B. S. Toguebaye
Laboratoire de Parasitologie Générale,
Département de Biologie Animale,
Faculté des Sciences et Technologies,
Université Cheikh Anta Diop, Dakar, Senegal

regionalization (L. Lavie et al., in preparation). A region having a biased GC content and consisting of blocks of repeated sequences, called the “subterminal region”, extends between the chromosome core and each subtelomere, just upstream of the rDNA unit. These blocks contain putative coding sequences (CDSs) that were generally annotated as “hypothetical proteins” in the *E. cuniculi* genome database. Several potential multigenic families were distinguished.

Shown to represent a transition area between euchromatin and heterochromatin in *Plasmodium falciparum* (Ralph and Scherf 2005), the subterminal region provides a particular niche on eukaryotic linear chromosomes and offers the possibility to define a particular gene context. In such a context, multigenic families encoding surface glycoproteins are often related to the pathogenicity of parasitic organisms (Barry et al. 2003). The process of antigenic variation in major protozoan parasites clearly involves surface proteins, such as VSGs in *Trypanosoma brucei* (Lalor et al. 1984) and PfEMP1 in *P. falciparum* (Baruch et al. 1995), with a strict control of the expression of different genes switching from one locus to another during infestation. In fungal species, multigenic families that are clustered in subterminal regions may encode adhesins, e.g. FLO and EPA families in *Saccharomyces cerevisiae* and *Candida glabrata*, respectively (Halme et al. 2004; Castano et al. 2005).

The microsporidian spore is surrounded by a highly resistant cell wall and uses a sophisticated mechanism to infect a potential host cell. A long polar tube is extruded at the anterior pole, the sporoplasm flows through this tube and is pulled into the host cell. Some *E. cuniculi* proteins have been clearly assigned to either the spore wall (Bohne et al. 2000; Peuvrel-Fanget et al. 2006) or polar tube (Delbac et al. 1998, 2001; Peuvrel et al. 2002). It is noteworthy that these proteins are encoded by single-copy genes located on chromosome cores. Taking into account the critical role of multigenic families in the virulence of other parasites, studies on such families in *E. cuniculi* may be of interest for a better understanding of host cell-microsporidia interactions. In this work, we investigated a set of repeated CDSs that are scattered on the consensus sequences of most chromosome extremities, called the *interB* family. Following bioinformatic analysis, PCR primers were designed to amplify the whole gene repertoire. Reverse transcription (RT)-PCR experiments provided evidence for the transcription of some gene copies. We demonstrated that *interB*-like genes are present in two other *Encephalitozoon* species (*E. intestinalis* and *E. hellem*) and two species from different genera (*Brachiola algerae* and *Vittaforma corneae*), all sharing the capacity to infect humans.

Materials and methods

Parasite isolates and cell culture

Encephalitozoon cuniculi, *E. hellem* and *E. intestinalis* species were grown *in vitro* in either Madin-Darby canine kidney (MDCK) or Human Foreskin Fibroblasts (HFF) cells at 37°C as described elsewhere (Beauvais et al. 1994). Twelve different *Encephalitozoon* isolates were used: four isolates from strains I and III of *E. cuniculi* (EcG serving as genome reference, Katinka et al. 2001; CH-H9, Mathis et al. 1997; MX-H5, Mathis et al. 1997 and CDC:V282, Croppo et al. 1997), four isolates from two genotypes of *E. hellem* (CDC:V213, Xiao et al. 2001; CDC:V242, Xiao et al. 2001; CDC:V257, Schwartz et al. 1993 and EhD also called “Didier isolate”, Biderre et al. 1999a) and four from *E. intestinalis* (CDC:V316; CDC:V307, del Aguila et al. 1998 and EvG also called “Van Gool isolate”, Biderre et al. 1999a). These isolates are part of the laboratory collection. They were provided by collaborators and maintained in culture in our laboratory. *Brachiola algerae* cells were propagated on HFF cell layers at 31°C (isolate provided by the Institute for Medical Parasitology, Bonn, Germany, Müller et al. 2000).

Spores collected from supernatants were purified as described previously (Delbac et al. 1998) and stored at 4°C in PBS until use. *Vittaforma corneae* spores were kindly provided by Dr E.S. Didier (Tulane National Primate Research Center, USA) as methanol fixed material until DNA extraction. The DNA of the other species was isolated from spores extracted from host using protocols based on Percoll gradients (Jouvenaz 1981). Fish microsporidia were purified from wild animals: *Glugea atherinae* from *Atherina boyeri* collected in the south of France (Biderre et al. 1994), *Spraguea lophii* from anglerfish (Mansour et al. 2004) and *Neonosemoides syacii* from *Syacium micrurum* Turbot brought to fishermen in the region of Dakar (Faye et al. 1994). Insect microsporidia were purified from bred animals: silkworm for *Nosema bombycis* [the production from silk gland was performed in China, Southwest University, Chongqing: Sato and Watanabe (1980) protocol] and *Gryllus bimaculatus* for *Paranosema grylli* (Nassonova et al. 2005).

PCR and RT-PCR amplification

Sequences of specific and generic *interB* primers are provided in Table 1. DNA extractions were performed from purified spores using an Elu-Quick kit (Schleicher and Schuell Bioscience) and DNA amplifications were carried out with a Perkin-Elmer GenAmp 2400

Table 1 Primer sequences

Primer	Sequence	Annealing temperature (°C)
Bsub1	TATARTATAGAAAGACTG	44
Bsub2	AAAAGAGAAGATGCGAAG	50
Bsub3	ARTCTTTYCGTGATACAG	50
Bsub4	GAAAACYCRCCTGCTGGA	56
Bsub5	TRTRWCAAGGTATGGAAG	50
Bsub6	TAAGGAACGCGCAGGGTT	56
newECU	GAGTCTTTTATGATAC	44
B2	GTCADGCAGATGCTGCTGGGAAG	76
GAR	RCGTGCWCCTCCAAYCTC	58
YHL	ACMTCAGGTGGTACAMAAC	56
FQI	CYYCCACACTAYYTGRAA	52
C1	CACCAGGTTGATTCTGGC	56
C2	GTGACGGGCGGTGTGTAC	60

Annealing temperature was calculated with $2(A + T) + 4(G + C)$ method for small primers and with nearest-neighbour formula (salt concentration at 50 mM) for B2

thermal cycler. PCR reactions were done in 50- μ l total volume with 1 ng of DNA and 2.5 U of Taq DNA polymerase (Eurobio). Samples were subjected to 30 cycles of amplification, first denaturing DNA for 30 s at 94°C, then annealing for 1 min (temperature between 50 and 62°C depending on the primers used), and elongation at 72°C for 1 min. A final extension for 15 min at 72°C was used.

Total RNA of *E. cuniculi*-infected cells were extracted using the RNeasy® Mini Kit (Qiagen). Samples were then treated with RNase-free DNase I (Gibco) and complete elimination of DNA was verified by a direct PCR assay including the same primers as for RT-PCR. Synthesis of cDNA was performed with 0.5 μ g of total RNA and the reverse primer FQI, using manufacturer recommendations for RNA LA PCR Kit (Takara).

All PCR products were cloned in pGEM®-T Easy Vector (Promega) and sequenced using the ABI Prism Dye Terminator Cycle Sequencing kit according to the recommendations of the manufacturer (Perkin-Elmer). Direct and reverse reads were overlapping. A consensus was built for each clone. Sequence redundancy was observed for most *interB* loci. Clones were redundant when originating from the same isolate. The redundancy was measured in the whole clone insert at the DNA level. Non-redundant sequences have been deposited in the database (EMBL accession numbers AM295101 to AM295152). Corresponding newECU-GAR clones had no strict homology to chromosome IV 3' extremity. NewECU02/NewECU04 specific B2-YHL clones resulting from transcriptional analysis gave only one type of sequence. As some divergence was only found with the last 100 bp of the chromosome consensus, a bad quality of the assemblage in this region seemed likely. Sequences provided by the different experiments were assembled to produce the New-

ECU02_EcG_ENCCU consensus, representative of the two new *interB* loci.

Sequence analysis and multiple sequence comparison

Search for sequence similarities was carried out using the Basic Local Alignment Search Tool (Blastn) from NCBI BLAST web portal. Protein motifs and domains were predicted using the following web services: Pfam database at Sanger Institute or INRA portals, SignalP at CBS web portal, TMPred at Swiss EMBNET node, and secondary structure prediction algorithms (SOPMA, DSC and PHD) implemented at NPSA web site (<http://www.npsa-pbil.ibcp.fr>). The sequence multiple comparison and trees were generated by the neighbour joining (NJ) method implemented in CLUSTALW 1.83 software available for Windows XP. Multiple alignments were edited using Bioedit v7.0.5 for Windows XP. The tree plots were produced by the njplot program.

Pulsed field gel electrophoresis

Intact DNA was prepared in plugs from 10^8 purified spores embedded in 0.7% low-melting-point agarose (Cambrex Bioscience). Agarose plugs were incubated for 2 h at 50°C in a first lysis buffer consisting of 200 mM EDTA (ethylene diamine tetraacetic acid), 100 mM DTT (dithiothreitol) and 5% SDS. After rinsing in distilled water, they were incubated for 48 h at 50°C in a second lysis buffer containing 500 mM EDTA, 0.5% *N*-lauryl sarcosine and 2 mg/ml proteinase K. Plugs were stored in 500 mM EDTA at 4°C until gel electrophoresis.

Pulsed field gel electrophoresis (PFGE) separation of chromosomal DNA molecules from *B. algerae* spores was performed at 12°C in 1% agarose (Seakem GTG,

Cambrex Bioscience) gel with 0.5× TBE buffer (45 mM Tris-HCl, pH 7.5, 45 mM boric acid and 1 mM EDTA). Different electrophoreses were used to characterize the karyotype. For *E. hellem*, agarose concentration was increased to 1.2% and PFGE was carried out for 36 h with 16-s pulse at 200 V. Lambda phage DNA concatemers and *Saccharomyces cerevisiae* chromosomes (New England Biolabs) were used for DNA size estimation. After electrophoresis, gels were stained in 0.5 µg/ml ethidium bromide solution.

Two-dimensional (2-D) genome analysis involving the digestion of the whole molecular karyotype was performed by the KARD (karyotype and restriction display) 2-D PFGE method (Brugère et al. 2000b). *B. algerae*- and *E. hellem*-KARD profiles were obtained using *Mlu*I and *Bss*HII enzymes (New England Biolabs) as rare cutters. Pulse program in the second dimension step was 5-s pulse for 12 h and 15-s pulse for 8 h with *E. hellem* DNA. Various three-step pulse programs were used with *B. algerae*. Migrations were performed at a voltage of 200 V in 0.5× TBE buffer (10°C). The agarose gel concentration was 1.2%. Prior to hybridization, KARD gels were dried as previously described (Brugère et al. 2000b).

Probes and hybridization assays

Hybridizations were performed on dried gels and nylon membranes. Southern blot analyses of karyotype DNA were performed using the Hybond-N+ membrane (GE Healthcare). DNA molecules were partially depurinated in 0.125 M HCl, denatured in 1.5 M NaCl/0.5 M NaOH and neutralized in 1.5 M NaCl/0.5 M Tris-HCl before an overnight transfer in 20× SSC (0.3 M Na citrate, 3 M NaCl). Membranes were cooked at 80°C for 2 h to fix DNA.

The rDNA specific probe was a 1,200-bp *ssu* gene fragment amplified with C1 and C2 primers (Table 1). After amplification, PCR products were cloned in pGEM-T vector (Promega) and sequenced. Vector-free specific probes were obtained by insert amplification using universal T7 and SP6 primers. DNA was radiolabelled with [α -P32] dCTP by the random priming method using Klenow fragment of DNA polymerase I (GE Healthcare). After a prehybridization step of 1 h at 65°C, hybridizations were performed at 65°C overnight, in 5× SSC, 0.5% SDS and 5× Denhardt's (0.1% Bovine Serum Albumin, 0.1% Ficoll, 0.1% Polyvinylpyrrolidone). Membranes or gels were washed 2–3 times in 2× SSC, 0.1% SDS buffer at room temperature. Stringent washing steps were often performed for 15–30 min in 0.1× SSC, 0.1% SDS. Autoradiography was done using HyperfilmTM MP (GE Healthcare) and

two intensifying screens at −80°C. Long exposure times (3–4 days) were chosen to better visualize faint signals.

Results

Bioinformatic analysis of *E. cuniculi* *interB* family

Gene copies of the *E. cuniculi* *interB* family are located on nearly all chromosome extremities, the apparent exceptions being on the 5' side of chromosomes III and V consensus and on the 3' side of chromosome X consensus (Table 2). The redundancy of potential InterB proteins is moderate, with 25 CDSs for 15 different protein sequences (15 Uniprot entries). Protein size varies from 579 to 641 amino acid (aa) residues (average: 600 aa). No sequence homology was found in public databases.

The general organization of InterB sequences, predicted with bioinformatic tools, is illustrated in Fig. 1. Two transmembrane (TM) domains, 16 to 21 aa in length, may be present in the N-terminal region (positions close to 50 and 80). However, corresponding TMPred scores were rather low (less than 1,500). In most cases, only one of the two TM domains was considered to be significant (score close to 1,000), the other domain having a score below 500. An additional TM domain extending over the first 20 aa of some InterB members (Y111_ENCCU, Y205_ENCCU, Y208_ENCCU, Y4G7_ENCCU) was predicted to be a signal peptide. The N-terminal region contained two highly conserved peptide sequences: LMGL(L/F)AEG (labelled as B1) and V(K/M/R)QMLLGK (labelled as B2) at the average positions 250 and 320, respectively. Data from multiple sequence alignments and predicted structural features of N-terminal regions were suggestive of six different subfamilies. Some segments of the N-terminal region indeed typified each subfamily, justifying their use for the design of specific PCR primers: Bsub1, Bsub2, Bsub3, Bsub4, Bsub5 and Bsub6 (Table 1). The central part of InterB proteins was characterized by a variable EK-rich sequence preserving α -helix secondary structures. Most insertion–deletion events (Indels) were associated with this core region (Fig. 2). In contrast, the C-terminal part was the most conserved in sequence, showing 70 to 95% identity over a length of 210 aa (vs 50% identity over 280 aa in the N-terminal part). A PfamA domain of unknown function (DUF1609) has been assigned to this region. Three highly conserved motifs (GAR, YHL and FQI) were of interest for PCR reverse primer design (Table 1).

Table 2 *Encephalitozoon cuniculi* *interB* repository structure

CDS name	Accession number	Chrom. position	Status	Group	B2-GAR	B2-YHL	FQI	Bsub1-GAR	Bsub2-GAR	Bsub3-GAR	Bsub4-GAR	Bsub5-GAR	Bsub6-GAR
ECU01_0080	Y108_ENCCU	18529–17534	P	g12	227	—	y	—	—	—	—	—	504
ECU01_0110	Y111_ENCCU	24645–22867		g6	318	700	y	—	628	581	—	—	—
ECU01_1500	Y111_ENCCU	185338–187116		g6	318	700	y	—	628	581	—	—	—
ECU01_1530	Y108_ENCCU	191454–192449	P	g12	227	—	y	—	—	—	—	—	504
newECU02	NA	<–737	P	g11	294	664	y	—	—	—	—	—	—
ECU02_0050	Y205_ENCCU	5448–7019		g8	288	670	y	—	598	551	—	—	—
ECU02_0080	Y208_ENCCU	10535–12274		g9	279	661	y	—	—	—	670	—	—
ECU02_1560	Y108_ENCCU	191403–192398	P	g12	227	—	y	—	—	—	—	—	504
newECU03	NA	193169–>		g10	297	679	y	—	—	—	688	—	—
ECU04_0090	Y108_ENCCU	14964–13969	P	g12	227	—	y	—	—	—	—	—	504
ECU04_1670	Y4G7_ENCCU	211107–212867		g8	288	670	y	—	598	551	—	—	—
newECU04	NA	217386–>	P	g11	294	664	y	—	—	—	—	—	—
ECU05_1680	Y5G8_ENCCU	207816–209654		g1	384	766	y	—	—	—	—	—	660
Additional	YB03_ENCCU	NA		g2	309	691	y	—	—	—	651	—	—
ECU06_0040/0050	Y604/	4779–6721		g3	330	712	y	—	—	—	672	—	—
	Y605_ENCCU												
ECU06_0080	Y608_ENCCU	8293–10035		g4	279	658	y	534	—	—	—	—	—
ECU06_1620	Y6G2_ENCCU	207962–209701		g5	279	661	y	534	—	—	—	—	—
ECU06_1670/1680	Y6G8/	213571–215516		g3	330	712	y	—	—	—	672	—	—
	Y6G7_ENCCU												
ECU07_0050	Y705_ENCCU	5899–5249	Pt	g14	—	—	y	—	—	—	—	—	—
ECU07_0070	Y707_ENCCU	7406–9253		g7	288	667	y	—	—	551	—	—	—
ECU07_1880/1890	Y718/	223959–225882		g2	309	691	y	—	—	—	651	—	—
	Y719_ENCCU												
ECU08_0010	Y108_ENCCU	1743–748	P	g12	227	—	y	—	—	—	—	—	504
ECU08_0040	Y111_ENCCU	7859–6081		g6	318	700	y	—	628	581	—	—	—
ECU08_2090	Y108_ENCCU	230369–231364	P	g12	227	—	y	—	—	—	—	—	504
ECU09_0010	Q8STY9_ENCCU	<–1095	P	g13	197	—	NA	—	—	—	538	—	—
ECU09_2020	Y705_ENCCU	248027–248677	Pt	g14	—	—	y	—	—	—	—	—	—
ECU10_0010/0020	Y718/	2618–695		g2	309	691	y	—	—	—	651	—	—
	Y719_ENCCU												
ECU11_0030	YB03_ENCCU	4681–6606		g2	309	691	y	—	—	—	651	—	—
ECU11_0050	Y5G8_ENCCU	10011–8173		g1	384	766	y	—	—	—	—	—	660
Additional	Y718/	NA		g2	309	691	y	—	—	—	651	—	—
	Y719_ENCCU												

Uniprot peptide sequence accession number is given for each CDS. Some *interB* loci are partially present at the extremity of chromosome consensus. The missing coordinate is marked by > or <. Pseudogenes are labelled with the letters P (Status column). Two ORFs for the truncated protein Y705_ENCCU are marked with Pt. Outside the six subfamilies, 14 *interB* gene identity groups can be identified. Grouping is performed independently of the gene length, but at high identity threshold. Four groups correspond to pseudo-genes (g11, g12, g13 and g14). When a mismatch disrupts an *interB* sequence, the two CDSs are given (if available). This concerns genes presenting single G indel mutations. In the case of ECU02_0050, annotation gave no CDS in the ORF containing three prime gene extremities. The size prediction of amplicons was estimated using genome data and perfect matches with primers (Table 1). Additional amplicons obtained with mismatches are indicated in italics. In the case of g13, B2-GAR amplicon was obtained despite that B2 presents three mismatches. At the opposite, RT-PCR experiments (Fig. 4) confirmed that YHL does not match with g12 CDS sequences whereas the primer presents only one mismatch with genome sequence. Predicted negative PCR results are marked by “–”. No B2 sequences were detected for group g14 as CDSs correspond to truncated genes lacking the 5' region. The YHL site is highly degenerated at a locus corresponding to g13 sequence. The existence of FQI sequence was validated experimentally for the genes of groups g10 and g11 NA not available

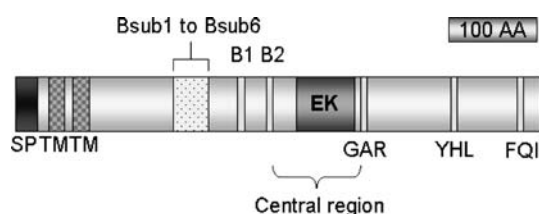


Fig. 1 Sequence features of putative InterB proteins in *E. cuniculi*. A signal peptide (SP) is predicted in subfamilies 2 and 5. One of the two potential transmembrane domains (TM) may have a significant score. Subfamily-specific motifs are around the position 230 (Bsub1 to Bsub6 designate eponym-specific primers). Conserved motifs used to design generic primers are indicated as B1 and B2 (forward orientation) and as GAR, YHL and FQI (reverse orientation). Size polymorphism is associated with the EK-rich region

Some putative InterB proteins were strongly truncated. This was the case of Y108_ENCCU (six CDSs) which corresponded to the N-terminal part of Y5G8_ENCCU (two CDSs). The truncation was due

to multiple transitions causing premature stop codons and to insertions introducing frame-shifts. At the 5' extremity of chromosome IX, mutations were spread all over an *interB* gene sequence, justifying the detection of only one small CDS (ECU09_0010). Peptide sequences deduced from different reading frames were homologous to the central and C-terminal parts of Y608_ENCCU. Three motifs (Bsub4, GAR and degenerated B2) were identifiable. The ECU09_0010 locus therefore appears as a chimera. Recombination events may also be hypothesized for two CDSs on two different chromosomes (ECU07_0050 and ECU09_2020). These CDSs encoded the same putative reduced protein (Y705_ENCCU), matching with only the C-terminal region of a typical InterB member. Single mutations (insertion/deletion of one G residue in a poly-G sequence) producing frame-shifts may justify four other truncated proteins: an Y4G7_ENCCU C-terminal variant (Y205_ENCCU), and three

Tran	E	E	K	A/T	K	S	
01	GAGGAGAAGGCAAGAGCAAGAAGGGAGGAAGAAA						AAGAGTGTGGGAGTCTCAG
02	GAGGAGAAGGCAAGAGCCAGAAGGGAGGAAGAAA						AAGAGTGTGGGAGTCTCAG
03	GAGGAGAAGGCAAGGGCAGGAAGGAT						GGGAAGAAGAAGAGTGTGAATGTTTCTG
04	GAGGAGAAGGCAAGAGCAAGAAAGAG						GGGGAAGAGGAAGAGTGTGGGAGTCTCCG
05	GAGGAGAAGGCAAGAGCAAGAAAGAG						GGGGAAGAGGAAGAGTGTGGGAGTCTCCG
06	GAGGAGAAGGCAAAAAGCAAGAAGAA						GGGGAAGAAGAAGAGTGTGGGAGTCTCTG
07	GAGGAGAAGGCAAGAGCAAGGAGAAGGCAAGAGCAAGGAGAAGGCAAAAAGCAAGAAGAAA						GGG-AGAGTGTGGGAGTCTCTG
08	GAGGAGAAGGCAAGAGCAAGGAGAAGGCAAGAGCAAGGAGAAGGCAAAAAGCAAGAAGAAA						GGG-AGAGTGTGGGAGTCTCTG
09	GAGGAGAAGGCAAGAGCAAGGAGAAGGCAAGAGCAAGGAGAAGGCAAAAAGCAAGAAGAAA						GGG-AGAGTGTGGGAGTCTCTG
10	(GAGGAGAAGGCAAGAGTAAAG) 2						AGGGGGAGGAAGGGGAAGAGTG-----CTTCTG
11	(GAGGAGAAGGCAAGAGTAAAG) 2						AGGGGGAGGAAGGGGAAGAGTG-----CTTCTG
12	(GAGGAGAAGGCAAGAGTAAAG) 4						AGGGGGAGGAAGGGGAAGAGTG-----CTTCTG
13	(GAGGAGAAGGCAAGAGTAAAG) 2						AGGGGGAG-AGGGGAAGAGTG-----CTTCTG
14	(GAGGAGAAGGCAAGAGTAAAG) 3						AGGGGGAG-AGGGGAAGAGTG-----CTTCTG
15	GAGGAGAAGGCAAGAGCAAGAAGA						GGGGGAAGAGGAAGAGTGTGGGAGTCTCCG
16	GAGGAGAAGGCAAGAGCAAGAAGA						GGGGGAAGAGGAAGAGTGTGGGAGTCTCCG
17	GAGGAGAAGGCAAGAGTAAAGAAAGGGAAGAAGAAAAGAAAAG						GAGTGCAGGAGTCCCTG
18	GAGGAGAAGGCAAGAGTAAAGAAAGGGAAGAAGAAAAGAAAAG						GAGTGCAGGAGTCCCTG
19	GAGGAGAAGGCAAGAGCAAGAAGAAAGGGAAGAAGAAGAAGAGAGAAA						GGGAGTCTCTG
20	GAGGAGAAGGCAAGAGCAAGAAGAAAGGGAAGAAGAAGAAGAGAGAAA						GGGAGTCTCTG
21	GAGGAGAAGCAAGAGCAAG						AAGAGTG-GGGAGTCTCTG
22	GAGGAGAAGCAAGAGCAAG						AAGAGTG-GGGAGTCTCTG
23	GAGGAGAAGCAAGAGCAAG						AAGAGTG-GGGAGTCTCTG
24	GAGGAGAAGCAAGAGCAAG						AAGAGTG-GGGAGTCTCTG
25	GAGGAGAAGCAAGAGCAAG						AAGAGTG-GGGAGTCTCTG
26	GAGAGAA (GGA) 3 (GAA) 3 (GGA) 3 2GAAGAGGAAGGAGGA (GAA) 3 (GAAGAAGGAGGA) 2 (GAA) 3 (GAAGAAGGAGGA) 8						GAAGAGTG-GGAAGTCTCTG
27	GAGAGAA (GGA) 3 (GAA) 3 (GGA) 3 2GAAGAGGAAGGAGGA (GAA) 3 (GAAGAAGGAGGA) 2 (GAA) 3 (GAAGAAGGAGGA) 8						GAAGAGTG-GGAAGTCTCTG

Fig. 2 Multiple sequence alignment for the most variable part in the central region of *interB* genes. Conserved pyrimidine positions are highlighted in grey. The beginning of the aligned sequences encodes a conserved motif (*top line*) despite that Indels are still detected in the upstream region. In the subfamily 6, the two bottom sequences (26 and 27) diverge from the others (21 to 25), the conserved motif being replaced by a long degenerated microsatellite (GPuPu)_n. In all subfamilies, this variable region is purine-rich and should encode a glutamic acid–lysine peptide able to form an alpha helix secondary structure. The alignment has been stopped at a pyrimidine-rich stretch for uncharged amino acids (leucine, valine, proline, serine). Frame shifts are at the origin of amino acid sequence variation at this position. The differences between subfamilies support an evolution based on duplication (e.g. Bsub4 genes) and pyrimidine–pyrimidine or purine–purine transitions (e.g. Bsub1 genes). Sequence order: 01, Bsub1_Y6G2_ENCCU-ECU06_1620; 02, Bsub1_Y608_ENCCU-ECU06_0080; 03, Bsub1-YHL_clo4_140104_V242_ENCHe;

04, Bsub3_Y707_ENCCU-ECU07_0070; 05, Bsub2-3_Y4G7_ENCCU-ECU04_1670; 06, B2-YHL_clo3_081205_EhD_ENCHe; 07, Bsub2-3_Y111_ENCCU-ECU01_0110; 08, Bsub2-GAR_clo16_030404_V242_ENCHe; 09, Bsub3-GAR_clo3_081204_VITCO; 10, Bsub4_YB03_ENCCU-ECU11_0030; 11, Bsub4-GAR_clo21_081205_EhD_ENCHe; 12, B2-GAR_clo6_081204_EhD_ENCHe; 13, B2-YHL_clo5_050105_BRAAL; 14, B2-YHL_clo8_050105_BRAAL; 15, Bsub5_Y208_ENCCU-ECU02_0080; 16, Bsub5-GAR_clo3_270404_V307_ENCIN; 17, Bsub5_newECU03_EcG_ENCCU; 18, B2-YHL_clo13_060205_BRAAL; 19, newECU02_EcG_ENCCU; 20, B2-YHL_clo11_170304_BRAAL; 21, Bsub6_Y108_ENCCU-ECU01_0080; 22, Bsub6-GAR_clo27_121205_EhD_ENCHe; 23, Bsub6-GAR_clo31_051212_EvG_ENCIN; 24, Bsub6-GAR_clo9_050705_BRAAL; 25, Bsub6-GAR_clo5_131204_VITCO; 26, Bsub6_Y5G8_ENCCU-ECU11_0050; 27, Bsub6-GAR_clo37_051212_EvG_ENCIN

YB03_ENCCU N-terminal variants (Y719_ENCCU, Y604_ENCCU and Y6G8_ENCCU). It seems likely that all the above-mentioned sequences may be representative of pseudogenes (marked with “P” in Table 2).

Thorough bioinformatic analyses of the *E. cuniculi* genome sequence enabled us to predict additional copies of existing genes (Table 2) from sequence homologies over long distance (several kbp) between chromosomes V (3') and XI (5') or X (5') and XI (3'). Blast homologies also revealed some partial *interB* genes that were not annotated as CDSs, given their very extreme position at one end of the consensus sequences for chromosomes II, III and IV. An experimental search for possible lacking members of the *interB* repertoire was therefore undertaken.

PCR amplification of *E. cuniculi interB* genes

The whole sequences of the three expected additional *interB* genes, called *newECU02*, *newECU03* and *newECU04*, were successfully trapped by PCR using extremity-specific primers in combination with conserved reverse oligonucleotides. The *newECU03* gene was Bsub5 specific. The design of an additional primer, called newECU (Table 1), allowed for the characterization of *newECU02* and *newECU04*. The NewECU02_EcG_ENCCU consensus for two loci was further built using B2-YHL clones. One mismatch was detected in a polyG tract close to the GAR motif, suggesting that these loci may represent pseudogenes. In spite of sequence homologies with subfamilies 2 and 3, they could not be identified with our Bsub primers. The NewECU-YHL consensus shared 98% identity over 235 bp with the two shortened CDSs representative of Y705_ENCCU protein. Thus, *newECU02* and *newECU04* may be viewed as equivalent to the complete forms of pseudogenes that were previously truncated. The total number of *interB* loci was finally estimated to be 30.

Taking advantage of the length polymorphism of the central gene region, the generic couple of primers B2-GAR were tested to amplify the core region of all *E. cuniculi interB* loci (Fig. 2). Agarose gel analysis of PCR products revealed four major bands, ranging from 200 to 400 bp (Fig. 3) and fitting with the expected size profile. Some pseudogenes gave small amplicons, which is consistent with the conception that these loci may have derived from recombination events preventing the reconstruction of complete gene copies. The diversity of *interB* genes was then investigated through PCR amplification with subfamily-specific forward primers coupled to the reverse primer GAR (Fig. 3).

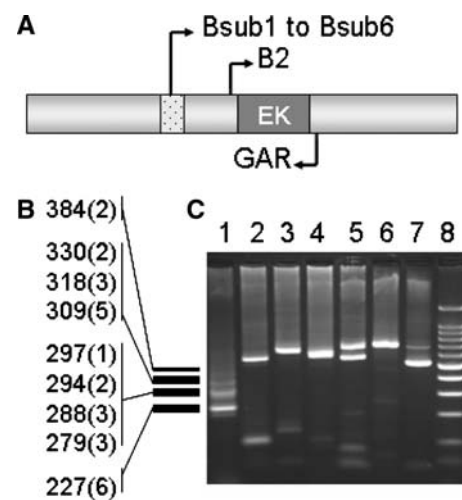


Fig. 3 PCR amplification of *E. cuniculi interB* sequences. **a** PCR strategy involving specific (Bsub1 to Bsub6, dotted motif) and generic primers (B2 and GAR). The procedure takes advantage of the size polymorphism in the central region (EK motif). **b** Expected B2-GAR electrophoretic profile of DNA fragments. *In silico* PCR was performed on genome sequence without mismatches. The sizes of different DNA fragments are indicated in base pairs on the left, with the number of copies per genome in brackets (see also Table 2). **c** Agarose gel analysis of PCR products. The obtained profile is in agreement with the predicted one for generic PCR B2-GAR (lane 1) as well as for PCR combining GAR reverse primer with specific forward oligonucleotides (lanes 2–8). Predicted size and intensity ratio between amplicates can be deduced from Table 2. Lane 2, Bsub1, one 534 bp fragment. Lane 3, Bsub2, two bands in a nearly equimolar ratio. Lane 4, Bsub3, two closely migrating bands. Lane 5, Bsub4, two bands. Lane 6, Bsub5, one band for two co-migrating fragments. Lane 7, Bsub6, two bands in a 1:3 ratio. Lane 8, 100-bp DNA ladder (Promega). A faint 200-bp band in the B2-GAR PCR profile (lane 1) corresponds to the amplification of a 197-bp fragment from a pseudogene on chromosome IX (locus ECU09_0010), in spite of three mismatches existing between B2 primer and genome sequence

PCR products were cloned and at least eight clones from each amplicon were sequenced. The obtained data validated the specificity of Bsub primers. The B2-GAR profiles of different *E. cuniculi* isolates were identical. Only CDC:V282 isolate showed a very limited sequence polymorphism (one base, one amino acid), when compared with the isolate used for genome sequencing.

RT-PCR detection of *interB* transcripts

The efficiency of the transcription of *E. cuniculi interB* genes was evaluated by RT-PCR amplification (Fig. 4a). The use of the B2-YHL primer couple should exclude truncated loci. Indeed, the YHL primer does not recognize pseudogene sequences forming the g12 (CDSs for Y108_ENCCU) and g13 (ECU09_0010) similarity groups, and B2 motif is absent in the g14

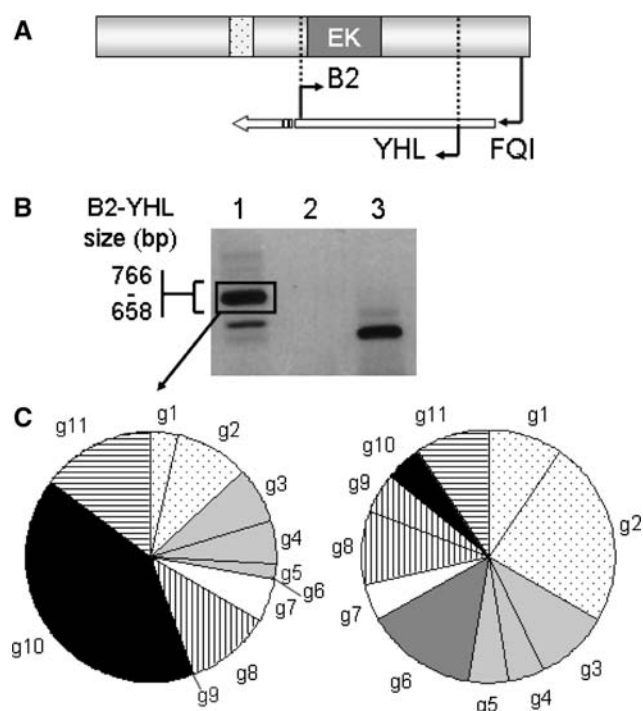


Fig. 4 Identification of *E. cucullis* *interB* transcripts. **a** RT-PCR strategy. FQI primer was used for cDNA synthesis and PCR amplification of all potential cDNAs was performed with B2 and YHL primers. Subfamily-specific motif (dotted box) and variable central region (EK box) are figured in *interB* gene schema. **b** Agarose gel analysis. Lane 1, RT-PCR assay with microsporidian RNA. A wide band, centred on 690 bp, fits with the expected size range. A lower band results of a PCR artefact, as shown by sequencing. Lane 2, DNA contamination control, without reverse transcription. Lane 3, host cell RNA control (RNA from non-infected cells). Cloning and sequencing of the products revealed no identity with *interB* family. **c** Comparison of the distribution of *interB* transcripts (on the left) with the distribution of corresponding genes in the *E. cucullis* genome (on the right). Transcripts were identified after cloning of B2-YHL PCR products, random selection of colonies and insert sequencing. The sequences fall into similarity groups g1 to g11 (Table 2). The predominant transcript was from g10 (*newECU03*)

group (ECU07_0050 and ECU09_2020). To test the absence of a skew due to the choice of oligonucleotides, the same PCR conditions were first applied to genomic DNA. After cloning of PCR products (Fig. 4b), bacterial colonies were randomly selected then inserts were sequenced and the distribution of corresponding sequences was compared with that of *interB* loci in the genome (Fig. 4c). As expected, no truncated pseudogenes were amplified and full-length loci were amplified with the same efficiency. RT-PCR experiments were performed with whole RNA extracted from two independent cultures. A total of 54 amplicons (38 in experiment 1 and 16 in experiment 2) were sequenced then assigned to similarity groups. In both experiments, the transcripts originated from

several *interB* genes on various chromosomal ends, but with a strong predominance (40–50%) of the transcript corresponding to the *newECU03* gene on chromosome III. Accordingly, the B2-GAR migration pattern exhibited a major band at about 290 bp (data not shown).

Evidence of *interB*-like genes in some other microsporidian species

The existence of two other *Encephalitozoon* species (*E. intestinalis* and *E. hellem*), well known to infect immune-deficient humans, offered the opportunity to test an intra-genus conservation of *interB* sequences. The PCR amplification strategy used for *E. cucullis* was applied to four *E. intestinalis* isolates (CDC:V297, CDC:V316, CDC:V307 and EvG) and four *E. hellem* isolates representative of two distinct genotypes (Wainwright-like CDC:V213 isolate versus CDC:V242, CDC:V257 and EhD isolates). Amplicons were successfully obtained with the B2-GAR couple of primers and their major size range (200–450 bp) was similar to *E. cucullis* one (Fig. 5). A striking additional feature was the presence of small-sized PCR products that may correspond to truncated loci. The total number of bands was sometimes reduced. For example, two major bands were found with *E. hellem* CDC:V242. These differences appear to be mainly of quantitative nature because further use of specific primers revealed the presence of all subfamilies in most isolates. The only exception was *E. hellem* EhD that was negative for amplification with Bsub2 and Bsub3 primers.

The search for *interB*-like genes was then extended to microsporidian species having different host specificity and belonging to seven different genera: *Vittaforma corneae* (a human parasite), *B. algerae* (an insect parasite found to also infect humans; Coyle et al. 2004),

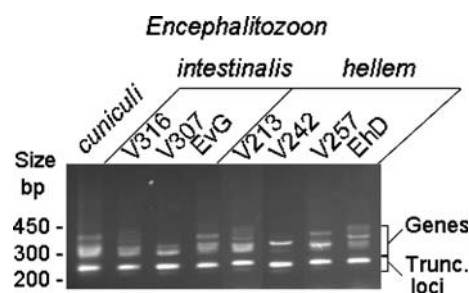


Fig. 5 Agarose gel analysis of PCR-amplified *interB* genes in *Encephalitozoon* species. PCR was performed using generic primers B2 and GAR. Three *E. intestinalis* isolates and four *E. hellem* isolates were compared with *E. cucullis* reference isolate. The observed bands are within the expected size range (200–400 bp). The smallest fragments correspond to truncated loci and are present in all isolates. None of the profiles are completely identical

Nosema bombycis and *Paranosema grylli* (two strict insect parasites), *Glugea atherinae*, *Neonosemoides syacii* and *Spraguea lophii* (three strict marine fish parasites). Amplicons were detected in only *V. corneae* and *B. algerae* (Fig. 6). Subsequent amplifications with *E. cuniculi* specific primers supported a repartition among the four following subfamilies: Bsub1, Bsub2, Bsub3 and Bsub6. In *B. algerae*, the B2-GAR profile was restricted to three bands (200, 250 and 300 bp) and Bsub6 was the only subfamily to be successfully amplified.

Interspecific comparison of *interB* sequences

The highest efforts for cloning and sequencing PCR products were performed in *E. hellem* EhD and *B. algerae*. No new *interB* subfamily was revealed. Analysis of generic clones (B2-GAR or B2-YHL PCR) only showed divergent genes within existing subfamilies. The sequences of the Bsub regions in these loci were too divergent to be detected by specific PCR but their clustering with *E. cuniculi* genes was unambiguous (Fig. 7). One sequence from *E. hellem* EhD fall into the g8 group (cluster Bsub2–3_Y4G7_ENCCU) and some clones of *B. algerae* fitted with subfamilies 4 and 5. Some other sequences from the two species clustered with the subfamily characterized by two new *E. cuniculi* genes (*newECU02* and *newECU04*). Moreover, *B. algerae* showed homologues of *newECU03*, the locus corresponding to the major transcript in *E. cuniculi*. The subfamilies 1, 2 and 3 were not repre-

sented in *B. algerae*. Sequence differences within a subfamily were often due to a single base mutation leading to one amino acid change. Indels were again detected in the central part of the genes (Fig. 2).

The phylogenetic tree for all identified *interB* sequences supports a subdivision into three major clusters (Fig. 7). Two clusters each include a single subfamily (Bsub4 or Bsub6) whilst a super-cluster is formed by the other subfamilies. The Bsub6 cluster is especially stable, as judged by the high redundancy of identical sequences. A gene with a conserved frame-shift in the central region, corresponding to the lowest band in B2_GAR profiles, is specifically associated with this cluster. One clone arisen from the highest band of the Bsub6-GAR profile in *E. intestinalis* EvG showed a long CT microsatellite insertion (160 bp) but the failure in the identification of a same clone from other isolates suggests a very rare case (Fig. 2). The variability of the Bsub4 cluster members is marked by a variable number (2–4) of repetitions of a 21-bp AG-rich sequence in their central part (Fig. 2). In spite of the lack of amplification with the Bsub4 specific primer, the DNA of *B. algerae* provided some B2-YHL clones that also carry Bsub4-type repetitions (clones 5, 6 and 8 of the 05/01/05). Amplifiable with four different primers (Bsub1, Bsub2, Bsub3 and Bsub5), the sequences in the third and largest *interB* cluster are characterized by a high rate of evolution. The subfamilies 2 and 3 largely diverge in *Encephalitozoon* species. The rapid evolution is also evident at the Indel level and frame-shifts over short distances account for frequent changes at the amino-acid level.

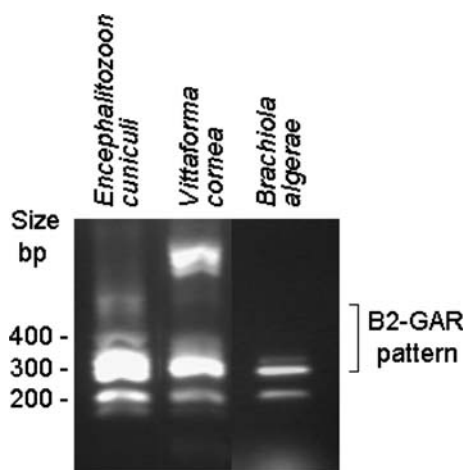


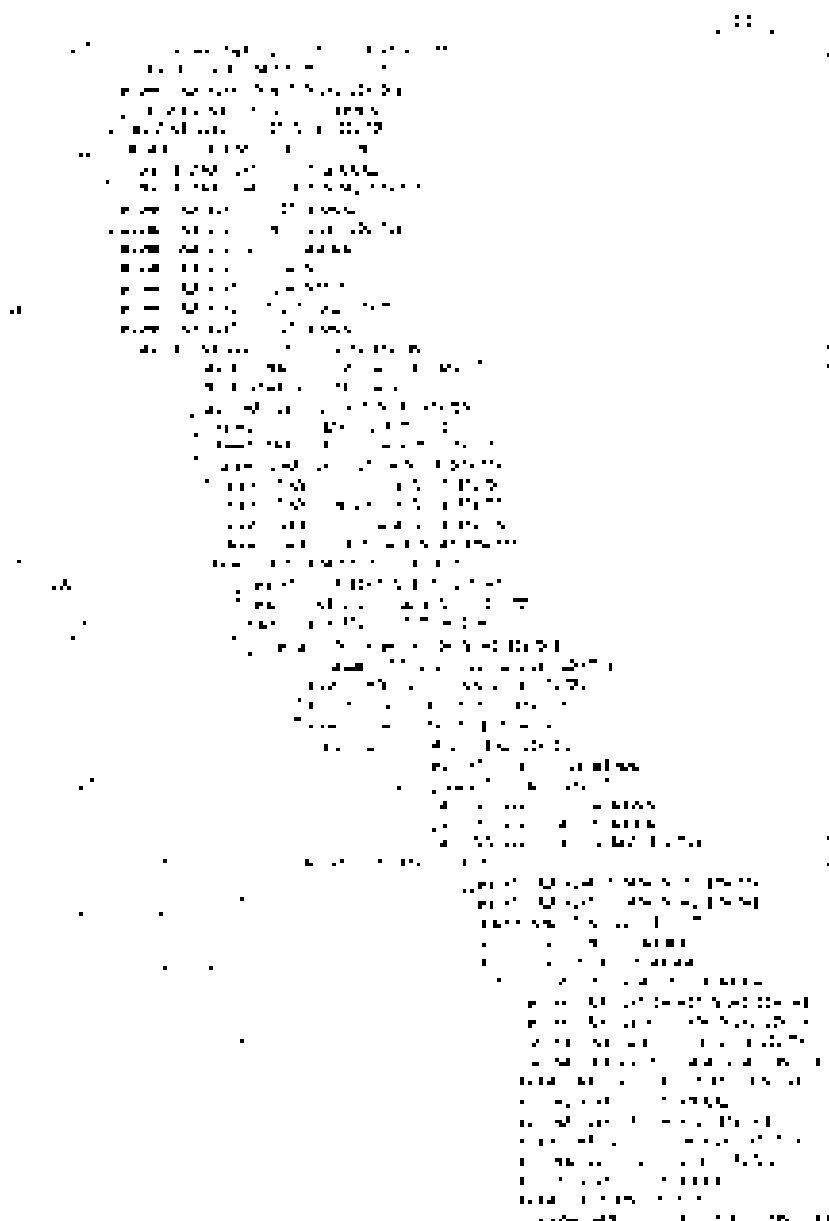
Fig. 6 Banding patterns of B2-GAR PCR products from *Vittaforma corneae* and *Brachiola algerae*. Compared with *E. cuniculi*, *V. corneae* exhibits a high proportion of unspecific and larger fragments (more than 400 bp) while *B. algerae* shows a simpler profile with three bands in the expected size range. In these two species, the only PCR product related to an *E. cuniculi* pseudo-gene belongs to the g12 group (227 bp)

Chromosomal distribution of *interB*-like genes in *E. hellem* and *B. algerae*

Following 1-D and 2-D PFGE analyses of *E. hellem* and *B. algerae* genomes, the chromosomal distribution of *interB*-like genes was studied by hybridization with generic (B2-YHL PCR product) and specific (from clones) probes. A possible linkage with rDNA genes was tested by hybridization with a small-subunit (*ssu*) rDNA probe. The use of a 2-D PFGE method (KARD) facilitated the determination of the relationship between karyotype bands and individual DNA molecules, after in-gel digestion of the whole karyotype with rare cutters (*Bss*HII and *Mlu*I) and comparison of the restriction patterns obtained in the second-dimension step (Brugère et al. 2000a; Cornillot et al. 2002). This method was also useful for differentiating some homologous chromosomes in diploid genomes.

The 1-D PFGE profile of *E. hellem* EhD consisted of ten bands (K1 to K10, Fig. 8). The upper band

Fig. 7 ClustalW tree for *interB* genes from different microsporidia based on the comparison of non-redundant central sequences. Three major clusters are revealed, as indicated on the *right* of the figure. Sequences were produced from inserts that were obtained after cloning of B2-GAR or Bsub-specific primer-GAR amplification products or less frequently of YHL PCR. Longer branches were observed when comparing peptide sequences (data not shown). Bootstrap values larger than 500 are given for longer branches. The sequence naming system is as follows: primer pair used for PCR_clone number_key date in lab. book_isolate name_species abbreviation. Specific primers, Uniprot accession numbers and CDS name for *interB* loci of the *E. cuniculi* reference isolate are also given. A consensus sequence was used for the NewECU02/new-ECU04 pseudogenes (New-ECU02_EcG_ENCCU)



(330 kbp) was not visualized in a previous study (Biderre et al. 1999a). The existence of this additional chromosome was confirmed by *MluI*-KARD analysis (see *MluI* restriction fragments, Fig. 8). This analysis revealed 11 different restriction profiles for the ten karyotype bands. A same number of restriction profiles were obtained with *BssHII* enzyme (data not shown). Thus, the *E. hellem* genome might comprise 11 different chromosomes, like in *E. cuniculi*. The sum of chromosome sizes provided 2.75 Mpb, as a new estimate of the haploid genome size. Having no evidence for length polymorphism of possible homologous chromosomes, we cannot currently decide whether the *E. hellem* genome is either haploid or diploid. All the

karyotype bands hybridized with *ssu* and *interB* probes (Fig. 8), suggesting that the subterminal location of *interB* genes is conserved in another member of the *Encephalitozoon* genus.

The molecular karyotype of *B. algerae* was characterized by a high number of bands ranging in size between 180 and 2,580 kbp (Fig. 9a). Somewhat like in *Trypanosoma* (Melville et al. 1998), three size categories may be distinguished: mega-base, intermediate and low. This complex microsporidian genome was subjected to KARD analysis using *MluI* enzyme, the average cutting frequency (one site per 200 kbp) of this enzyme being expected to be adequate for the digestion of most karyotype bands (Fig. 9b). The largest

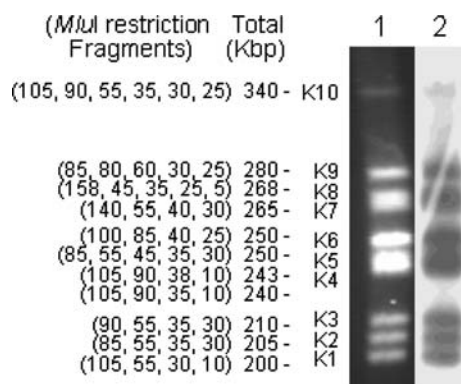


Fig. 8 One-dimensional PFGE analysis of *Encephalitozoon hellem* (EhD) genome and hybridization with *ssu* rDNA and *interB* (B2-YHL PCR) probes. Ten bands (K1–K10) are resolved from 200 to 340 kbp (lane 1). KARD-*MluI* analysis has revealed that K6 contains two different DNA molecules (data not shown). On the left, *MluI* restriction fragment sizes are therefore given for a total of 11 chromosomes. Both *interB* and *ssu* rDNA probes hybridize with all karyotype bands (lane 2, *interB* blot)

chromosome gave 12 restriction fragments (700, 500, 320 × 2, 160, 140, 110, 90, 70 × 2, 60, and 40 kbp). Twelve different DNA molecules were clearly assigned to the seven bands of the mega-base region. Our estimate of 27 DNA molecules in the intermediate region is more prone to error, because of the persistence of a high number of co-migrating entities in the second dimension. The 12 “mini-chromosomes”, identified in the low region, frequently contained only one *MluI* restriction site. Careful analysis of the *MluI*-KARD profiles of intermediate and mega-base regions supports a diploid status. The sum of all individual DNA sizes attains about 38 Mbp. Admitting that some chromosomes are twice represented, the haploid genome size of *B. algerae* should be between 15 and 20 Mbp.

Hybridization with *ssu* probe indicated that 15 *B. algerae* chromosomes harbour rDNA copies (Fig. 9c). Each of these chromosomes produced three reactive *MluI* fragments, including a common 20-kbp fragment. As the *ssu* probe overlaps a *MluI* restriction site, the 3' part of the probe may recognize a fragment carrying a large-subunit (*lsu*) rDNA gene. Thus, the 20-kbp fragment should represent the most extreme chromosome region, extending from one *lsu* rDNA gene to the telomere. Of variable size and generally large, the two other hybridizing fragments should contain a 5' part of the *ssu* rDNA gene associated with a variable part of the chromosome core. This is consistent with the location of a single rDNA unit on each extremity of the 15 considered chromosomes. The *interB* probe was found to hybridize with three rDNA-containing chromosomes. The two *interB* signals observed on the *MluI* profile of each of these chromosomes fitted with the

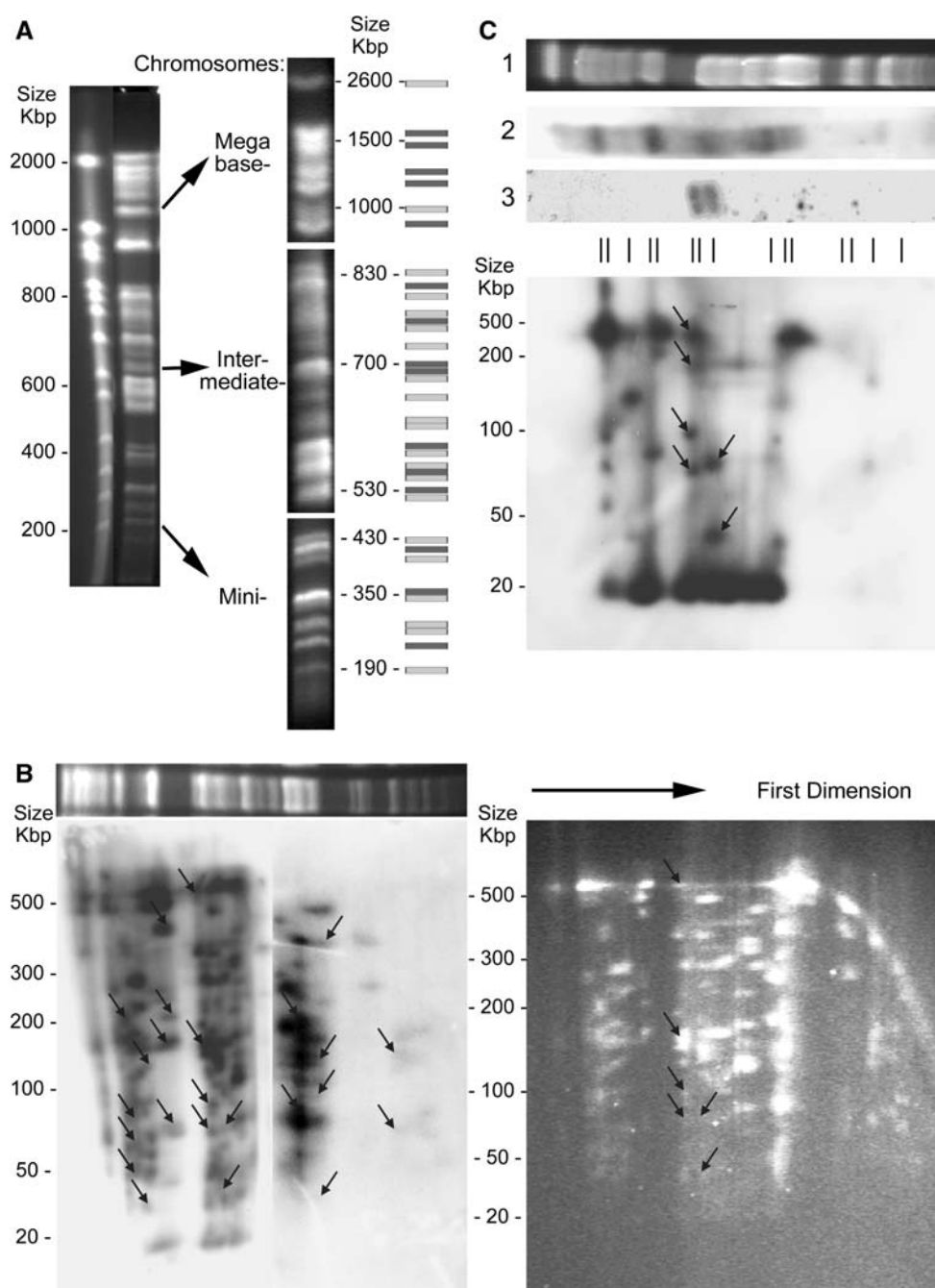
two variable *ssu* rDNA-hybridizing fragments (arrows in Fig. 9c). This indicates that the *interB* repertoire is confined to the rDNA-to-core transitional region in *B. algerae*, as in *E. cuniculi*.

Discussion

Chromosome location and expression of *interB* genes

Following PCR experiments, the *interB* multigenic family in the microsporidian *E. cuniculi* has been found to comprise a total of 30 members, all located in near-rDNA subterminal regions. Large gene families occupying subterminal regions in various organisms are the consequence of ectopic recombinations (Pryde et al. 1997). Some recombination events may justify chromosome length polymorphism (CLP), as described in *P. falciparum* (Corcoran et al. 1988) or *Pneumocystis carinii* (Cornillot et al. 2002). CLP related to Indels in chromosome extremities also exist in *E. cuniculi* (Biderre et al. 1999b; Brugère et al. 2000c, 2001) but the identical B2-GAR patterns in all tested *E. cuniculi* isolates suggest that DNA gains or losses are rather rare in *interB* areas. This may not be the case in *E. intestinalis* and *E. hellem* isolates that display different B2-GAR profiles. Comparative studies on the *ptp1* locus have also shown a reduced polymorphism in *E. cuniculi*, perhaps because of a more recent origin of this species (Peuvel et al. 2000; Xiao et al. 2001). Several facts argue for a close relationship between the *interB* gene context and euchromatin–heterochromatin junction. First, in both *Encephalitozoon* and *Brachiola*, the *interB* genes extend close to rRNA genes which are typical heterochromatin-specific genes. Second, these genes belong to blocks of sequences having a GC content higher than in all other *E. cuniculi* chromosome regions. Third, while the members of the fast-evolving cluster are often closer to the chromosome core, those exhibiting an extremely low inter-species polymorphism are closer to the chromosome ends that should be more protected from mutagens through a tight binding of the DNA to proteins in heterochromatin areas.

Protein expression from a multigenic family is often related to a process of allelic exclusion. For example, the active gene of the *vsg* repertoire in *T. brucei* can be expressed in one of about 20 “bloodstream form expression sites” that appear as hot spots of recombination (Becker et al. 2004). Although recombination and antigenic variation are closely linked mechanisms, the nature of the mono-allelic control remains unclear (Pays 2005). The expressed gene may differ in individuals having a same genetic background and in daughter



cells after mitosis. Our RT-PCR analysis of *interB* transcription in *E. cuniculi*-infected host cells has revealed a preferential activity of a gene copy positioned on one extremity of the chromosome III (*newECU03*) but we cannot affirm that a mono-allelic control occurs. It should be stressed that *E. cuniculi* develops asynchronously inside parasitophorous vacuoles, with accumulation of the latest sporogonial stages (immature and mature spores) in the old vacuoles. Differences in transcriptional activity between parasite stages may therefore justify an unequal distribution of different *interB* transcripts in the whole parasite population. A working

hypothesis is that the transcription of a set of *interB* genes would occur early in proliferative stages (meronts) and would be followed by an epigenetically controlled extinction process during sporogony. Noticeably, the epigenetic control of *var* gene expression in *P. falciparum* has been related to a moving junction between euchromatin and heterochromatin (Freitas-Junior et al. 2005).

The cellular localization and functional significance of InterB proteins are still unknown. The prediction of a signal peptide in some of these proteins, including the protein corresponding to the preponderant transcript,

Fig. 9 Two-dimensional PGFE KARD analysis of *Brachiola algerae* genome and hybridization with *ssu* and *interB* probe. **a** Separation of chromosomal DNAs in the first dimension step. *Left* Yeast PFGE marker and general overview of the *B. algerae* karyotype (60-s pulses for 15 h, 85-s pulses for 15 h and 140-s pulses for 15 h at 150 V). *Right* Three-step pulse programs were applied to improve the resolution of bands in different karyotype regions: “mega-base” region (60-s pulses for 8 h, 85-s pulses for 18 h and 160-s pulses for 19 h at 150 V), “intermediate” region (60-s pulses for 15 h, 85-s pulses for 15 h and 140-s pulses for 15 h at 150 V) and “mini-chromosomes” region (60-s pulses for 15 h, 85-s pulses for 15 h and 140-s pulses for 10 h at 150 V, the gel concentration was enhanced to 1.2%). The number of DNA molecules per karyotype band was estimated on the basis of fluorescence intensity and of the comparison of each band size with the sum of corresponding *Mlu*I fragment sizes. The molecular karyotype is schematized using *light grey boxes* for a single DNA molecule and *dark grey boxes* for two DNA molecules. **b** Autoradiography and ethidium bromide-stained pattern of restriction fragments after *Mlu*I digestion of the whole karyotype. *Top part left*: KARD first dimension. *Bottom left*: *Mlu*I-KARD autoradiographs. Four fragments per DNA molecule are generally produced in the intermediate region. Small fragments at

20 kbp are detected only after autoradiography. *Arrows* indicate fragments hybridizing with *ssu* probe (main signals in **c**). This section part is composed of two images, the *left* one being more resolute for large restriction fragments (8-s pulses for 8 h, 15-s pulses for 7 h and 40-s pulses for 7 h at 200 V), the *right* one being performed in the same conditions as the ethidium bromide stained gel (*right*). *Right part*: The number of *Mlu*I-restriction fragments is underestimated when KARD analysis is based on gel DNA fluorescence after ethidium bromide staining (pulse program: 5-s pulses for 8 h, 15-s pulses for 7 h and 30-s pulses for 9 h at 200 V). *Arrows* indicate fragments hybridizing with *interB* probe. **c** *B. algerae* *interB* repertoire genome organization. *Top part*: KARD first dimension and hybridizations. Lane 1, ethidium bromide-stained gel (60-s pulses for 15 h, 100-s pulses for 15 h and 160-s pulses for 17 h at 150 V). Lane 2, *ssu* rDNA hybridization. Lane 3, *interB* hybridization. *Bottom part*: Hybridization of *ssu* rDNA probe to *Mlu*I-KARD profile (pulse program: 5-s pulses for 8 h, 15-s pulses for 7 h and 30-s pulses for 7 h at 200 V). Reactive fragments derive from 15 chromosomes, their relative positions being indicated by *vertical lines*. Faint signals are observed after over-exposure of the film. *InterB*-hybridizing fragments (*arrows*) are found on three chromosomes. The same first dimension was used for the gel presented at the right of **b**

suggests a possible translocation into the endoplasmic reticulum. As at least one TM domain is predicted in the N-terminal region, a targeting to a membrane system may also be assumed. A possible candidate is the plasma membrane of proliferative stages. However, the hypothesis of an export to the host cell surface via interactions with host cell trafficking machinery also deserves the attention to justify some aspects of the cellular immunity response to microsporidian infections (Moretto et al. 2004). An answer is expected from further investigations using immunocytochemical procedures with specific antibodies raised against a recombinant protein derived from the expression of the *newECU03* gene copy in a heterologous system.

Existence of an *interB* repertoire in human-infecting microsporidia

Over 10 microsporidian species, three *Encephalitozoon* species and two representatives of the *Vittaforma* and *Brachiola* genera, were *interB*-positive and harboured multiple gene sequences. Specific PCRs and sequence alignments have confirmed that the *interB* genes are dispersed in more than one cluster. As the acquisition of such genes via a lateral gene transfer process seems unlikely, the *interB* repertoire should have been present in a common ancestor of *Encephalitozoon*, *Vittaforma* and *Brachiola*. Several cytological features argue for distantly related genera. *Encephalitozoon* spp. are mononucleate cells and are enclosed inside parasitophorous vacuoles whereas both *V. corneae* and *B. algerae* display two abutted nuclei (dikaryon) and develop in direct contact with the host

cytoplasm. In addition, we have shown that the chromosome distribution in *B. algerae* is more complex (three different chromosome size classes) than in *Encephalitozoon* (only minichromosomes) and resembles those in some *interB*-negative species such as *Nosema bombycis* (Kawakami et al. 1994) and *Glugea atherinae* (unpublished data). A recent comparative *ssu* rDNA analysis of 125 microsporidian species has distinguished five major clades that mainly group according to habitat and host, which led the authors to propose three habitat-based classes: Aquasporidia (clades I, II and V) in hosts of freshwater origin, Marinosporidia (clade III) in hosts of saltwater origin and Terresporidia (clade IV) in hosts of terrestrial origin (Vossbrinck and Debrunner-Vossbrinck 2005). A highly simplified tree in Fig. 10 illustrates the repartition of the species that were tested for *interB*. The five positive ones belong to two very distinct clades: clade IV for *Encephalitozoon* spp and *V. corneae*, clade V for *B. algerae*. Thus, a close evolutionary relationship between *Encephalitozoon*, *Vittaforma* and *Brachiola* is not supported by *ssu* rDNA phylogeny. Moreover, as the clade IV also comprises an *interB*-negative species (*N. bombycis*), it cannot be considered that all representatives of Terresporidia contain an *interB* repertoire. Following their exclusion from the *Nosema* genus through independent phylogenetic studies, the insect pathogens *Paranosema grylli* and *Antonospora locustae* in clade II represent two very closely related genera and possibly a same genus (Sokolova et al. 2003; Slamovits et al. 2004). Given that no *interB* genes were identified in *P. grylli* and no Blast records were obtained between *interB* genes and partial *A. locustae*

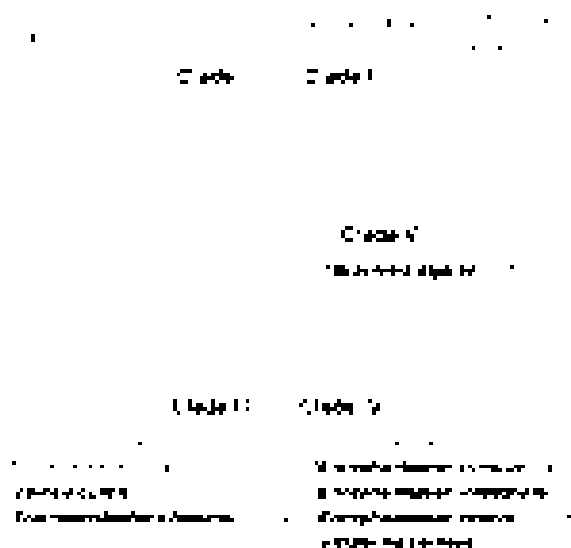


Fig. 10 Repartition of the studied species (names preceded by *asterisks*) between the five major clades of Microsporidia described by Vossbrinck and Debrunner-Vossbrinck (2005). Species infecting humans (in **bold**) vary in host specificity. A characteristic host order is indicated in *brackets*, except for *Vittiforma corneae* and *Trachipleistophora hominis* that have so far been isolated from only humans. *InterB*-positive species are *boxed*. Various *Amblyospora* species fall in clade I but remain to be investigated. For clade III, we mentioned *Vavraia culicis* and *T. hominis* to emphasize on cases that might be similar to *B. algerae* and *V. corneae*, respectively

genome sequence data (gmod.mbl.edu/antonospora), we assume that *A. locustae* is also *interB*-negative. The three studied marine fish parasites are *interB* negative, *Glugea* and *Spraguea* being assigned to the clade III. The phylogenetic position of *Neonosemoides* has not been investigated because of the lack of available rDNA sequence, but its habitat strongly suggests another member of Marinosporidia. A larger sampling of species is however required before concluding with the homogeneity of some clades. The history of the *interB* family during microsporidian evolution may have been dependent on duplication-divergence events and extensive chromosome rearrangements, leading to either expansion or loss of the family in species of different lineages. A comparison of the genome sequences of *E. cuniculi* and *A. locustae*, has suggested that the reduction of chromosome sizes is an ongoing process that takes place throughout the microsporidian phylum (Stechmann 2004).

The five *interB*-containing species have been recognized as human pathogens. However, as reviewed by Didier et al. (2004) and Mathis et al. (2005), a zoonotic potential likely exists for most human-infecting microsporidia. *E. cuniculi* is well known to have a wide range of mammalian hosts, with a high prevalence in rabbits.

E. intestinalis has been identified in gorillas whereas *E. hellem* can infect some birds. Although *V. corneae* has been detected only in humans so far, the *ssu* rDNA tree shows a clustering with *Cystospora* and *Endoreticulatus* species that invade lepidopterans. The ability to infect some poikilothermic animals is quite clear for *B. algerae* that was primarily recognized as a parasite of mosquitoes (Vavra and Undeen 1970). Both mosquito and human isolates of *B. algerae* can infect immunodeficient mice and can be grown on mammalian cell lines, even at the temperature of 37°C (Lowman et al. 2000; Cali et al. 2004; Kucerova et al. 2004; Visvesvara et al. 2005). Thus, some insects may be or may have been major sources of microsporidia for mammals. Although much remains to be learned about the relationship between genome size reduction and life cycle simplification in microsporidian groups, it is tempting to assume that the very small genomes in *Encephalitozoon* spp. are regressive forms of a *Brachiola*-like ancestral genome. Three obligatory blood-sucking mosquito genera (*Aedes*, *Anopheles* and *Culex*) are known as vectors of human diseases and can be infected by *B. algerae*. We propose an evolutionary scenario starting with the appearance of *interB* genes in an ancestor that was a parasite of mosquitoes. In the late Jurassic period, the evolution of mosquitoes and mammals offered the possibility for microsporidia to enter in contact with mammals. Then, a homoeothermic environment would have accelerated a regressive pathway driving the emergence of *Encephalitozoon* species.

Tracking *interB* genes in a higher number of microsporidian species should be useful to establish some correlations with the placement of rDNA units on chromosomes and with major features of the host-microsporidia relationship. In clade III, the case of *Trachipleistophora hominis* is of special interest for testing a specific origin of *interB* repertoire among mosquito parasites, because this human pathogen has been successfully transmitted to mosquito larvae and is closely related to insect-infecting *Vavraia* species (Weidner et al. 1999). In spite of the still limited knowledge of its distribution among microsporidia, the *interB* family may represent a set of valuable molecular markers for some human-infecting species and may offer perspectives of application in diagnostic and epidemiological fields.

Acknowledgments We thank all people who provided us with either DNA or spores from microsporidia that were not available in the Clermont-Fd laboratory: Drs Lamjed Mansour for *Glugea atherinae* and *Spraguea lophii*, Ngor Faye for *Neonosemoides syacii*, Elena Nassanova for *Paranosema grylli* and Wang Lin for *Nosema bombycis*. We convey our special thanks to Drs

Elisabeth S. Didier and Lisa C. Bowers for sending *Vittaforma corneae* spores despite the fact that these persons were affected by Katrina hurricane. We also thank Dr. Damien Brosson for helpful discussion. This work was supported by grants from EGIDE France (Ndongo Dia PhD) and the Medical Research Foundation (FRM: Fondation pour la Recherche Médicale; Laurence Lavie post-doctoral position).

References

- Barry JD, Ginger ML, Burton P, McCulloch R (2003) Why are parasite contingency genes often associated with telomeres? *Int J Parasitol* 33:29–45
- Baruch DI, Pasloske BL, Singh HB, Bi X, Ma XC, Feldman M, Taraschi TF, Howard RJ (1995) Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell* 82:77–87
- Beauvais B, Sarfati C, Challier S, Dérouin F (1994) *In vitro* model to assess effect of antimicrobial agents on *Encephalitozoon cuniculi*. *Antimicrob Agents Chemother* 38:2440–2448
- Becker M, Aitchison N, Byles E, Wickstead B, Louis E, Rudenko G (2004) Isolation of the repertoire of VSG expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast. *Genome Res* 14:2319–2329
- Bidre C, Pages M, Méténier G, David D, Bata J, Prensier G, Vivarès CP (1994) On small genomes in eukaryotic organisms: molecular karyotypes of two microsporidian species (Protozoa) parasites of vertebrates. *C R Acad Sci III* 317:399–404
- Bidre C, Canning E, Méténier G, Vivarès CP (1999a) Comparison of two isolates of *Encephalitozoon hellem* and *E. intestinalis* (Microspora) by pulsed field gel electrophoresis. *Eur J Protistol* 35:194–196
- Bidre C, Mathis A, Deplazes P, Weber R, Méténier G, Vivarès CP (1999b) Molecular karyotype diversity in the microsporidian *Encephalitozoon cuniculi*. *Parasitology* 118:439–445
- Bohne W, Ferguson DJ, Kohler K, Gross U (2000) Developmental expression of a tandemly repeated, glycine- and serine-rich spore wall protein in the microsporidian pathogen *Encephalitozoon cuniculi*. *Infect Immun* 68:2268–2275
- Brugère JF, Cornillot E, Méténier G, Bensimon A, Vivarès CP (2000a) *Encephalitozoon cuniculi* (Microspora) genome: physical map and evidence for telomere-associated rDNA units on all chromosomes. *Nucleic Acids Res* 28:2026–2033
- Brugère JF, Cornillot E, Méténier G, Vivarès CP (2000b) In-gel DNA radiolabelling and two-dimensional pulsed field gel electrophoresis procedures suitable for fingerprinting and mapping small eukaryotic genomes. *Nucleic Acids Res* 28:E48
- Brugère JF, Cornillot E, Méténier G, Vivarès CP (2000c) Occurrence of subtelomeric rearrangements in the genome of the microsporidian parasite *Encephalitozoon cuniculi*, as revealed by a new fingerprinting procedure based on two-dimensional pulsed field gel electrophoresis. *Electrophoresis* 21:2576–2581
- Brugère JF, Cornillot E, Bourbon T, Méténier G, Vivarès CP (2001) Inter-strain variability of insertion/deletion events in the *Encephalitozoon cuniculi* genome: a comparative KARD-PFGE analysis. *J Eukaryot Microbiol Suppl*:50S–55S
- Cali A, Weiss LM, Takvorian PM (2004) An analysis of the microsporidian genus *Brachiola*, with comparisons of human and insect isolates of *Brachiola algerae*. *J Eukaryot Microbiol* 51:678–685
- Castano I, Pan SJ, Zupancic M, Hennequin C, Dujon B, Cormack BP (2005) Telomere length control and transcriptional regulation of subtelomeric adhesins in *Candida glabrata*. *Mol Microbiol* 55:1246–1258
- Corcoran LM, Thompson JK, Walliker D, Kemp DJ (1988) Homologous recombination within subtelomeric repeat sequences generates chromosome size polymorphisms in *P. falciparum*. *Cell* 53:807–813
- Cornillot E, Keller B, Cushion MT, Méténier G, Vivarès CP (2002) Fine analysis of the *Pneumocystis carinii* f. sp. *carinii* genome by two-dimensional pulsed-field gel electrophoresis. *Gene* 293:87–95
- Coyle CM, Weiss LM, Rhodes LV 3rd, Cali A, Takvorian PM, Brown DF, Visvesvara GS, Xiao L, Naktin J, Young E, Garrec M, Colasante G, Wittner M (2004) Fatal myositis due to the microsporidian *Brachiola algerae*, a mosquito pathogen. *N Engl J Med* 351:42–47
- Croppo GP, Visvesvara GS, Leitch GJ, Wallace S, De Groote MA (1997) Western blot and immunofluorescence analysis of a human isolate of *Encephalitozoon cuniculi* established in culture from the urine of a patient with AIDS. *J Parasitol* 83:66–69
- del Aguila C, Croppo GP, Moura H, Da Silva AJ, Leitch GJ, Moss DM, Wallace S, Slemenda SB, Pieniazek NJ, Visvesvara GS (1998) Ultrastructure, immunofluorescence, western blot and PCR analysis of eight isolates of *Encephalitozoon (Septata) intestinalis* established in culture from sputum and urine samples and duodenal aspirates of five patients with AIDS. *J Clin Microbiol* 36:1201–1208
- Delbac F, Peyret P, Méténier G, David D, Danchin A, Vivarès CP (1998) On proteins of the microsporidian invasive apparatus: complete sequence of a polar tube protein of *Encephalitozoon cuniculi*. *Mol Microbiol* 29:825–834
- Delbac F, Peuvrel I, Méténier G, Peyretailade E, Vivarès CP (2001) Microsporidian invasion apparatus: identification of a novel polar tube protein and evidence for clustering of *ptp1* and *ptp2* genes in three *Encephalitozoon* species. *Infect Immun* 69:1016–1024
- Desportes I, Le Charpentier Y, Galian A, Bernard F, Cochand-Priollet B, Lavergne A, Ravisse P, Modigliani R (1985) Occurrence of a new microsporidian: *Enterocytozoon bien-eusi* n.g., n. sp., in the enterocytes of a human patient with AIDS. *J Protozool* 32:250–254
- Didier ES, Stovall ME, Green LC, Brindley PJ, Sestak K, Didier PJ (2004) Epidemiology of microsporidiosis: sources and modes of transmission. *Vet Parasitol* 126:145–166
- Faye N, Toguebaye BS, Bouix G (1994) *Nosemoides syacii* n.sp., a microsporidian parasite of the West African turbot *Syacium micrurum* Ranzani, 1840. *Syst parasitol* 29:43–50
- Freitas-Junior LH, Hernandez-Rivas R, Ralph SA, Montiel-Condado D, Ruvalcaba-Salazar OK, Rojas-Meza AP, Mancio-Silva L, Leal-Silvestre RJ, Montijo AM, Shorte S, Scherf A (2005) Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell* 121:25–36
- Halme A, Bumgarner S, Styles C, Fink GR (2004) Genetic and epigenetic regulation of the FLO gene family generates cell-surface variation in yeast. *Cell* 116:405–415
- Jouvenaz DP (1981) Percoll: an effective medium for cleaning microsporidian spores. *J Invertebr Pathol* 37:319
- Katinka MD, Duprat S, Cornillot E, Méténier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissbach J, Vivarès CP (2001) Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* 414:450–453

- Kawakami Y, Inoue T, Ito K, Kitamizu K, Hanawa C, Ando T, Iwano H, Ishihara R (1994) Identification of a chromosome harbouring the small subunit ribosomal RNA gene of *Nosema bombycis*. J Invertebr Pathol 64:147–148
- Kucerova Z, Moura H, Visvesvara GS, Leitch GJ (2004) Differences between *Brachiola (Nosema) algerae* isolates of human and insect origin when tested using an *in vitro* spore germination assay and a cultured cell infection assay. J Eukaryot Microbiol 51:339–343
- Lalor TM, Kjeldgaard M, Shimamoto GT, Strickler JE, Konigsberg WH, Richards FF (1984) Trypanosome variant-specific glycoproteins: a polygene protein family with multiple folding patterns? Proc Natl Acad Sci USA 81:998–1002
- Lowman PM, Takvorian PM, Cali A (2000) The effects of elevated temperatures and various time-temperature combinations on the development of *Brachiola (Nosema) algerae* N. Comb. in mammalian cell culture. J Eukaryot Microbiol 47:221–234
- Mansour L, Cheikali C, Desautais P, Coulon JP, Daubin J, Hasine OK, Vivares CP, Jeanjean J, Cornillot E (2004) Description of an ultrathin multiwire proportional chamber-based detector and application to the characterization of the *Spraguea lophii* (Microspora) two-dimensional genome fingerprint. Electrophoresis 25:3365–3377
- Mathis A, Michel M, Kuster H, Muller C, Weber R, Deplazes P (1997) Two *Encephalitozoon cuniculi* strains of human origin are infectious to rabbits. Parasitology 114:29–35
- Mathis A, Weber R, Deplazes P (2005) Zoonotic potential of the microsporidia. Clin Microbiol Rev 18:423–445
- Melville SE, Leech V, Gerrard CS, Tait A, Blackwell JM (1998) The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* and the assignment of chromosome markers. Mol Biochem Parasitol 94:155–173
- Moretto M, Weiss LM, Khan IA (2004) Induction of a rapid and strong antigen-specific intraepithelial lymphocyte response during oral *Encephalitozoon cuniculi* infection. J Immunol 172:4402–4409
- Müller A, Trammer T, Chioralia G, Seitz HM, Diehl V, Franzen C (2000) Ribosomal RNA of *Nosema algerae* and phylogenetic relationship to other microsporidia. Parasitol Res 86:18–23
- Nassonova E, Cornillot E, Méténier G, Agafonova N, Kudryavtsev B, Skarlato S, Vivarès CP (2005) Chromosomal composition of the genome in the monomorphic diplokaryotic microsporidium *Paranosema grylli*: analysis by two-dimensional pulsed-field gel electrophoresis. Folia Parasitol (Praha) 52:145–57
- Pays E (2005) Regulation of antigen gene expression in *Trypanosoma brucei*. Trends Parasitol 21:517–520
- Peuvel I, Delbac F, Méténier G, Peyret P, Vivarès CP (2000) Polymorphism of the gene encoding a major polar tube protein PTP1 in two microsporidia of the genus *Encephalitozoon*. Parasitology 121:581–587
- Peuvel I, Peyret P, Méténier G, Vivarès CP, Delbac F (2002) The microsporidian polar tube: evidence for a third polar tube protein (PTP3) in *Encephalitozoon cuniculi*. Mol Biochem Parasitol 122:69–80
- Peuvel-Fanget I, Polonais V, Brosseau D, Texier C, Kuhn L, Peyret P, Vivarès C, Delbac F (2006) EnP1 and EnP2, two proteins associated with the *Encephalitozoon cuniculi* endospore, the chitin-rich inner layer of the microsporidian spore wall. Int J Parasitol 36:309–318
- Pryde FE, Gorham HC, Louis EJ (1997) Chromosome ends: all the same under their caps. Curr Opin Genet Dev 7:822–828
- Ralph SA, Scherf A (2005) The epigenetic control of antigenic variation in *Plasmodium falciparum*. Curr Opin Microbiol 8:434–440
- Sato R, Watanabe H (1980) Purification of mature microsporidian spores by isodensity equilibrium centrifugation. Journal of Sericulture Science Jpn 49:512–516
- Schwartz DA, Visvesvara GS, Diesenhouse MC, Weber R, Font RL, Wilson LA, Corrent G, Serdarevic ON, Rosberger DF, Keenen PC, Grossniklaus HE, Hewan-Lowe K, Bryan RT (1993) Pathologic features and immunofluorescent antibody demonstration of ocular microsporidiosis (*Encephalitozoon hellem*) in seven patients with acquired immunodeficiency syndrome. Am J Ophthalmol 115:285–292
- Slamovits CH, Williams BA, Keeling PJ (2004) Transfer of *Nosema locustae* (Microsporidia) to *Antonospora locustae* n. comb. based on molecular and ultrastructural data. J Eukaryot Microbiol 51:207–213
- Sokolova YY, Dolgikh VV, Morzhina EV, Nassonova ES, Issi IV, Terry RS, Ironside JE, Smith JE, Vossbrinck CR (2003) Establishment of the new genus *Paranosema* based on the ultrastructure and molecular phylogeny of the type species *Paranosema grylli* Gen. Nov., Comb. Nov. (Sokolova, Selezniev, Dolgikh, Issi 1994), from the cricket *Gryllus bimaculatus* Deg. J Invertebr Pathol 84:159–172
- Stechmann A (2004) Genome evolution: the dynamics of static genomes. Curr Biol 14:473–474
- Vavra J, Undeen AH (1970) *Nosema algerae* n. sp. (Cnidospora, Microsporidia) a pathogen in a laboratory colony of *Anopheles stephensi* Liston (Diptera, Culicidae). J Protozool 17:240–249
- Visvesvara GS, Moura H, Leitch GJ, Schwartz DA, Xiao LX (2005) Public health importance of *Brachiola algerae* (Microsporidia)—an emerging pathogen of humans. Folia Parasitol (Praha) 52:83–94
- Vossbrinck CR, Debrunner-Vossbrinck BA (2005) Molecular phylogeny of the Microsporidia: ecological, ultrastructural and taxonomic considerations. Folia Parasitol (Praha) 52:131–142
- Weidner E, Canning EU, Rutledge CR, Meek CL (1999) Mosquito (Diptera: Culicidae) host compatibility and vector competency for the human myositic parasite *Trachipleistophora hominis* (Phylum Microspora). J Med Entomol 36:522–525
- Xiao L, Li L, Moura H, Sulaiman I, Lal AA, Gatti S, Scaglia M, Didier ES, Visvesvara GS (2001) Genotyping *Encephalitozoon hellem* isolates by analysis of the polar tube protein gene. J Clin Microbiol 39:2191–2196

Chapitre 3 : Expression des protéines InterB

I- Expression de protéines InterB recombinantes

Une stratégie d'expression hétérologue des protéines InterB d'*Encephalitozoon cuniculi* chez *Escherichia coli* et de purification des protéines recombinantes a été mise en place, en vue de produire des anticorps spécifiques pour détecter ces protéines chez le parasite.

Dans un premier temps, ce travail a porté sur une partie du gène *InterB* correspondant au CDS *ECU01_0110* qui code pour la protéine Y111_ENCCU. Ce gène appartient à la sous-famille 2. Le choix de cette protéine était motivé par sa forte homologie à des protéines de surface. La séquence choisie pour l'expression correspond à la partie N-terminale dépourvue de la séquence correspondant au peptide signal prédit sur la protéine (entre les positions 20 et 288 acide aminé).

Par la suite, compte-tenu des résultats des analyses du répertoire transcrit, le CDS *newECU03* correspondant au transcrit majoritaire a plus particulièrement retenu notre attention (**Figure 41**). Après réalisation d'alignements multiples des séquences *interB* et recherche d'épitopes, quatre régions ont été ciblées pour les amplifications PCR:

- ✓ Un fragment de 111 pb (37 aa) spécifique à *newECU03*, avec épitope, amplifié par le couple d'amorces 5newEcu03spec1 Vs 3newEcu03spec2.
- ✓ Un fragment de 90 pb (30 aa) spécifique à *newECU03*, sans épitope, amplifié par le couple d'amorces 5newEcu03spec1 Vs 3newEcu03spec3.
- ✓ Un fragment générique (936 pb, 312 aa), sans épitope, amplifié par le couple d'amorces 5newEcu03spec1 Vs 3newEcu03gener1.
- ✓ Un fragment central générique (156 pb, 52 aa), avec épitope, amplifié par le couple d'amorces 5newEcu03gener2 Vs 3newEcu03gener2.

Les séquences des différentes amorces utilisées pour les amplifications sont présentées dans le **Tableau 4**.

M	R	W	K	H	I	I	W	I	L	I	T	V	I	D	E	M	H	18
ATG	CGA	TGG	AAG	CAT	ATA	ATA	TGG	ATA	TTA	ATT	ACT	GTG	ATT	GAT	GAG	ATG	CAC	54
C	S	S	E	L	E	G	A	I	S	F	E	G	M	T	G	C	E	36
TGC	TCC	TCG	GAA	TTA	GAA	GGC	GCC	ATT	TCT	TTT	GAA	GGA	ATG	ACG	GGC	TGC	GAG	108
S	R	V	I	L	F	P	F	V	F	K	G	N	N	I	V	I	S	54
AGT	CGA	GTA	ATA	TTG	TTT	CCG	TTT	GTA	TTC	AAA	GGA	AAT	AAC	ATA	GTA	ATA	TCA	162
P	T	T	K	Y	N	D	L	K	R	N	T	R	D	T	V	M	G	72
CCA	ACA	ACA	AAA	TAT	AAT	GAT	TTG	AAG	AGA	AAC	ACT	AGA	GAT	ACA	GTT	ATG	GGT	216
F	V	S	E	I	S	F	M	M	W	Y	F	L	F	G	R	V	K	90
TTC	GTT	AGT	GAG	ATT	TCC	TTT	ATG	ATG	TGG	TAT	TTT	CTA	TTT	GGT	AGG	GTA	AAA	270
G	S	D	R	L	E	R	L	F	D	V	G	M	K	A	Y	L	E	108
GGT	AGT	GAC	CGA	CTT	GAA	AGG	TTG	TTT	GAT	GTA	GGT	ATG	AAA	GCA	TAC	TTG	GAG	324
D	I	S	I	G	T	S	K	V	W	M	N	G	N	K	T	F	S	126
GAC	ATA	TCT	ATA	GGT	ACA	TCA	AAG	GTG	TGG	ATG	AAT	GGA	AAC	AAG	ACT	TTT	AGT	378
E	L	L	E	L	V	Y	E	R	L	F	Q	C	N	E	R	E	G	144
GAA	TTG	TTG	GAA	TTG	GTT	TAT	GAA	AGG	CTA	TTC	CAA	TGT	AAT	GAG	AGG	GAG	GGT	432
G	H	V	T	R	Y	G	R	N	I	I	E	E	V	D	D	M	I	162
GGG	CAT	GTA	ACA	AGG	TAT	GGA	AGG	AAT	ATT	ATA	GAA	GAG	GTT	GAT	GAT	ATG	ATA	486
K	N	I	A	A	D	I	G	K	E	E	E	E	M	W	I	K	A	180
AAA	AAC	ATC	GCT	GCT	GAC	ATA	GGC	AAA	GAA	GAA	GAG	GAA	ATG	TGG	ATA	AAG	GCT	540
L	N	D	H	R	E	H	G	V	S	L	C	N	R	E	V	W	K	198
TTA	AAC	GAC	CAT	AGA	GAA	CAC	GGG	GTG	TCT	CTT	TGT	AAT	AGA	GAG	GTG	TGG	AAG	594
Q	V	V	S	V	E	K	I	V	C	N	I	C	K	R	I	C	I	216
CAG	GTA	GTC	AGT	GTT	GAA	AAG	ATA	GTC	TGC	AAT	ATA	TGC	AAA	AGA	ATC	TGT	ATT	648
D	L	K	E	E	E	L	L	G	L	M	A	E	G	S	V	R	K	234
GAT	CTT	AAG	GAA	GAA	GAG	TTA	CTA	GGA	TTA	ATG	GCT	GAG	GGG	AGT	GTG	AGA	AAG	702
I	L	K	A	K	F	G	E	E	K	V	S	D	M	S	Y	L	E	252
ATT	TTG	AAA	GCA	AAA	TTT	GGT	GAA	GAA	AAG	GTC	AGT	GAT	ATG	AGT	TAT	CTA	GAA	756
F	A	V	V	D	I	S	L	L	L	D	A	H	K	E	Y	G	E	270
TTT	GCC	GTT	GTT	GAT	ATC	AGT	CTA	TTA	TTA	GAT	GCA	CAC	AAA	GAG	TAT	GGA	GAA	810
E	V	M	K	E	L	V	K	Q	M	L	L	G	K	N	G	K	E	288
GAG	GTT	ATG	AAG	GAA	TTG	GTT	AAG	CAG	ATG	CTG	CTG	GGG	AAG	AAT	GGA	AAG	GAG	864
I	N	R	R	Y	V	D	K	V	A	G	V	V	R	E	I	Q	R	306
ATA	AAC	AGG	AGA	TAT	GTT	GAC	AAA	GTT	GCA	GGT	GTA	GTC	AGG	GAA	ATA	CAG	AGA	918
R	R	E	R	E	I	E	K	N	L	R	E	L	L	K	D	E	E	324
AGA	AGA	GAA	AGA	GAG	ATA	GAA	AAG	AAC	CTG	AGG	GAG	CTG	CTA	AAA	GAC	GAG	GAG	972
K	A	K	S	K	K	K	G	K	K	K	K	K	R	S	A	G	V	342
AAG	GCG	AAG	AGT	AAA	AAG	AAA	GGG	AAG	AAG	AAA	AAG	AAA	AGG	AGT	GCA	GGA	GTC	1026
P	E	S	K	E	E	E	K	K	E	S	E	T	E	E	V	E	A	360
CCT	GAG	AGT	AAG	GAA	GAG	GAG	AAG	AAG	GAG	AGC	GAG	ACC	GAG	GAA	GTT	GAG	GCA	1080
G	E	E	T	E	M	L	S	E	E	V	G	G	A	R	R	K	T	378
GGT	GAG	GAG	ACG	GAG	ATG	CTA	TCC	GAG	GAG	GTT	GGA	GGA	GCA	CGC	AGG	AAG	ACA	1134
G	K	K	S	E	G	G	R	K	R	Y	K	I	H	R	R	V	L	396
GGT	AAG	AAA	AGC	GAG	GGT	GGC	CGG	AAG	CGC	TAC	AAG	ATC	CAC	AGA	AGG	GTT	CTC	1188

R	W	R	K	S	P	E	K	I	K	E	E	W	D	K	G	S	E	414
CGG	TGG	AGA	AAG	AGC	CCA	GAG	AAG	ATC	AAG	GAG	GAG	TGG	GAC	AAA	GGA	AGC	GAA	1242
E	R	W	R	G	R	S	L	E	E	I	K	E	Q	K	V	F	H	432
GAG	AGG	TGG	AGG	GGA	AGG	TCT	CTT	GAA	GAG	ATC	AAG	GAG	CAG	AAG	GTG	TTC	CAT	1296
D	I	V	G	V	L	E	L	L	R	S	E	D	A	D	R	F	F	450
GAC	ATA	GTG	GGT	GTG	CTG	GAA	CTT	CTC	AGG	AGT	GAA	GAT	GCA	GAC	AGG	TTC	TTT	1350
I	D	T	G	E	H	S	K	G	G	S	E	R	Q	R	M	V	A	468
ATA	GAT	ACA	GGA	GAG	CAC	AGC	AAG	GGC	GGG	AGT	GAG	AGG	CAG	AGG	ATG	GTG	GCG	1404
I	G	V	L	E	S	G	E	K	K	M	P	G	V	V	E	V	G	486
ATT	GGG	GTT	CTT	GAG	AGC	GGA	GAG	AAG	AAG	ATG	CCA	GGA	GTG	GTG	GAG	GTT	GGG	1458
T	F	K	D	S	P	S	G	C	P	V	V	Y	H	L	M	F	R	504
ACG	TTC	AAG	GAC	TCT	CCT	AGT	GGA	TGC	CCA	GTT	GTG	TAC	CAC	CTG	ATG	TTC	CGG	1512
V	T	G	I	E	E	I	G	G	V	M	S	P	G	F	A	E	A	522
GTT	ACG	GGG	ATA	GAG	GAG	ATA	GGA	GGT	GTT	ATG	AGC	CCA	GGG	TTT	GCC	GAG	GCC	1566
N	D	I	E	K	I	D	E	D	K	E	Y	Q	D	A	G	K	F	540
AAT	GAT	ATT	GAG	AAG	ATA	GAT	GAG	GAT	AAG	GAA	TAC	CAG	GAT	GCG	GGC	AAG	TTT	1620
V	Y	P	K	G	V	R	F	E	T	V	K	G	T	D	V	F	Q	558
GTG	TAT	CCA	AAA	GGA	GTA	AGG	TTT	GAA	ACA	GTG	AAG	GGG	ACC	GAT	GTA	TTC	CAG	1674
I	V	W	G	N	P	S	D	T	S	E	V	F	Y	R	L	T	I	576
ATA	GTG	TGG	GGG	AAT	CCC	TCT	GAC	ACC	TCT	GAG	GTT	TTC	TAC	CGT	CTC	ACG	ATT	1728
R	C	R	P	C	V	I	*											584
CGG	TGC	AGG	CCA	TGT	GTG	ATC	TGA											1752

Figure 41: Séquence de la protéine InterB codée par le CDS newECU03.

Le peptide signal de 19 aa (en rouge) et les épitopes choisis (en marron et soulignés) ont été respectivement prédits avec les programmes SignalP et B-epitope.

Les amplifications PCR sont réalisées dans un volume total de 50 µl (avec 1 ng d'ADN de *E. cuniculi* et 2,5 U de Taq DNA polymérase, Eurobio) à l'aide d'un thermocycler Perkin-Elmer 2400. Après une première étape de dénaturation de 3 min à 94°C, 30 cycles d'amplification sont effectués selon les conditions suivantes : une dénaturation de 30 s à 94°C, suivie d'une hybridation de 30 s à une température choisie en fonction des T_m des amorces utilisées, puis une phase d'élongation d'une minute à 72°C. La phase d'élongation finale est de 15 min à 72°C.

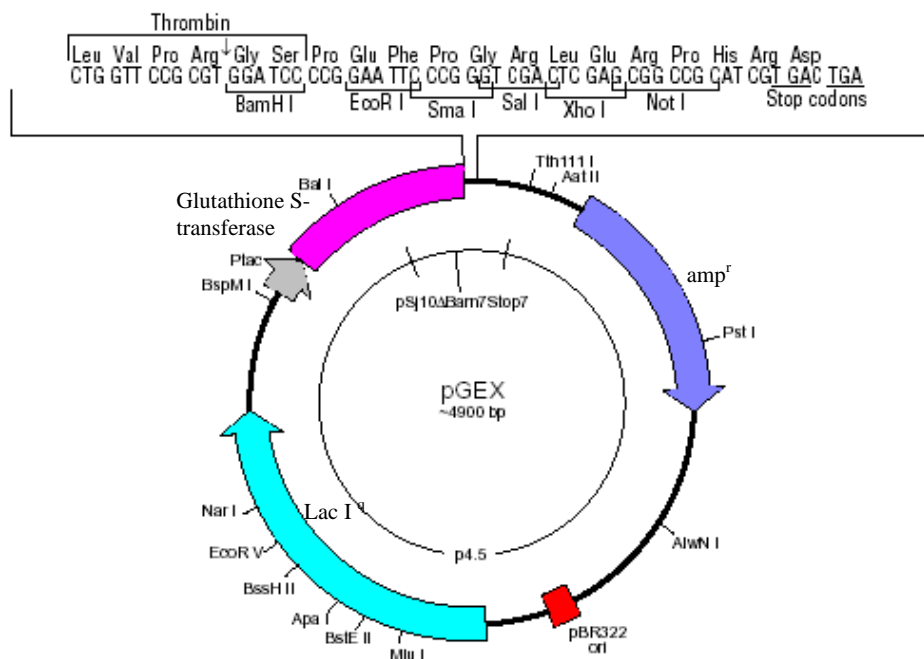
Afin de vérifier les tailles des fragments amplifiés, les produits PCR sont déposés avec un marqueur de taille (échelle 100 pb, Promega) sur un gel d'agarose à 1,2 %. Ce dernier, après migration, est traité au bromure d'éthidium (0,4 µg /ml) puis visualisé sous UV.

Amorces	Séquences	TM
ECU01_0110_f	CG GGATCC TCGGAAGAAGAGAATGC	74
ECU01_0110_r	GG GAATTC CTTCTTCCCCAGCAGCA	74
5newEcu03spec1	CG GGATCC AACACTAGAGATACAGTTATGGGTT	68
3newEcu03spec2	CG GAATTC TACTACAAACAACCTTTCAAGTCGG	70
3newEcu03spec3	CG GAATTC TCGGTCACTACCTTTTACCCTAC	68
3newEcu03gener1	CG GAATTC CCTGCGTGCTCCTCCAACCTC	70
5newEcu03gener2	CG GGATCC GAAAGAGAGATAGAAAAGAACCTGA	68
3newEcu03gener2	CG GAATTC TGCCTCAACTTCCTCGGTCTCGC	74

Tableau 4: Séquences et TM (températures de fusion) des amorces utilisées pour l'amplification de différents fragments de gènes *interB*. Les sites de restriction (***EcoRI*** GAATTC et ***BamHI*** GGATCC) utilisés pour le clonage sont indiqués en bleu gras.

I-1 Production de protéines recombinantes

Les bandes amplifiées sont découpées et purifiées sur colonne (Gel Extraction Kit, Qiagen). Les fragments sont par la suite digérés par les enzymes reconnaissant les sites de restriction introduits à leurs extrémités, puis mis en ligature avec le vecteur d'expression pGEX-4T1-His (**Figure 42**) préalablement coupé par les mêmes enzymes de restriction. En fait, il s'agit du vecteur pGEX-4T1 (Pharmacia) qui a été modifié par addition d'un épitope de 8 histidines en aval du site multiple de clonage. En amont du site, ce vecteur présente une séquence qui code la glutathion-S-transférase (GST). Ainsi, la protéine de fusion obtenue pourra être purifiée par chromatographie d'affinité sur colonne de nickel ou de glutathion.



Figures 42: Le vecteur d'expression *pGEX-4T1-His* (Pharmacia).

Le vecteur a été modifié par introduction d'une séquence codant pour un épitope de 8 résidus histidine entre les sites de restriction *XhoI* et *NotI*. Les fragments d'intérêt sont clonés entre les sites de restriction *BamHI* et *EcoRI*.

Après transformation et étalement sur milieu LB-ampicilline (100 µg/ml) de bactéries compétentes *E. coli* XL1-Blue, les clones sont criblés par restriction et vérifiés par séquençage (avec une amorce choisie en amont du site multiple de clonage).

L'expression a été effectuée chez des bactéries compétentes *E. coli* BL21+. Ces dernières sont transformées avec les plasmides recombinants puis étalées sur un milieu LB-ampicilline (100 µg/ml) et placées à 37°C pendant une nuit. Les transformants obtenus sont mis en culture dans 4 ml de milieu LB liquide contenant de l'ampicilline (100 µg/ml), sous agitation à 37°C jusqu'à obtention d'une densité optique de 0,5 (lecture à 600 nm). L'expression des protéines recombinantes est induite par addition d'IPTG (2 mM), pendant 4 h sous agitation. Les bactéries sont récupérées par centrifugation (5000 x g, 3 min), puis reprises dans un tampon de Laemmli à base de SDS 2,5 % - DTT 100 mM. Les protéines bactériennes ainsi solubilisées sont séparées sur gel SDS-PAGE à 10 % (Laemmli, 1970). La

révélation des bandes protéiques est effectuée au bleu de Coomassie à 0,2 % dans une solution de méthanol /acide acétique/eau distillée (4/2/4), pendant 30 min sous agitation à température ambiante. Les gels sont ensuite lavés dans une succession de bains du mélange méthanol /acide acétique/eau distillée (4/2/4).

I-2 Obtention des anticorps polyclonaux

Les bandes protéiques correspondant aux protéines recombinantes sont découpées, puis broyées au moyen d'un homogénéisateur de type Dounce dans 200 µl d'eau distillée stérile. Les broyats sont ensuite homogénéisés avec l'adjuvant de Freund complet (Sigma) puis injectés par voie intra-péritoneale à des souris (Swiss CD1, Charles River). Pour les rappels (14, 21, 28, 49 et 56 jours après la première injection), les antigènes sont mélangés à l'adjuvant de Freund incomplet. Le sang de l'animal est prélevé deux semaines après le dernier rappel par exsanguination totale et laissé coaguler 1 h à 37°C. Après une nuit à 4°C, le caillot rétracté est éliminé et le sérum centrifugé (20000 x g, 5 min) est conservé à - 20°C.

I-3 Extraction et séparation des protéines sporales

Des cellules HFF sont infestées avec des spores d'*E. cuniculi* souche GB-M1 (utilisée pour le séquençage). Après collecte des surnageants, les spores sont concentrées par centrifugation (3000 x g, 10 min). Les contaminants issus des cellules-hôtes sont éliminés par traitement au SDS 1% à 65°C, pendant 10 min, suivi de plusieurs lavages par centrifugation en présence de PBS.

L'extraction des protéines totales est réalisée par immersion des spores dans un tampon de lyse contenant SDS 0,2 %, DTT 100 mM et CHAPS (3-[(3-Cholamidopropyl)-Dimethylammonio]-1-Propane Sulfonate) 4 % et à l'aide de plusieurs cycles de congélation/décongélation dans de l'azote liquide suivi d'un traitement aux ultrasons (10 pulses de 30 s). Le matériel est repris dans une solution d'extraction contenant de l'urée 7M, thiourée 2M, DTT 100 mM, SDS 0,2 % et du CHAPS 4 % (6 h à température ambiante). Les protéines solubilisées sont collectées et conservées à -20°C.

Les échantillons protéiques (20 à 30 µg) sont séparés sur un gel de polyacrylamide à 10 ou 12 %. L'électrophorèse est effectuée sous intensité constante à raison de 25 mA par gel. Les gels sont colorés au bleu de Coomassie.

I-4 Transfert sur membrane et immunomarquage

Les protéines séparées en SDS-PAGE sont transférées sur membrane de PVDF (Poly Vinylidene DiFluoride, Polylabo) par un système de transfert semi-sec (Amersham).

Pour le marquage immunologique, une étape de saturation est effectuée avec du PBS-lait écrémé 5% pendant 1 h. La membrane est ensuite incubée pendant 3 h avec les anti-sérums dilués en PBS-Triton X-100 0,1 % (dilutions entre 1/200 et 1/500). Après 4 lavages (sous agitation) au PBS-Triton X-100 0,1 %, la membrane est traitée avec un anticorps secondaire anti-IgG de souris couplé à la phosphatase alcaline (Promega) dilué au 1/10000. Suite à 4 lavages en PBS-Triton X-100 0,1 % et un rinçage final en PBS seul, la révélation est faite en présence de NBT et BCIP selon les prescriptions du fournisseur (Promega).

I-5 Immunolocalisation

Des cellules HFF cultivées au contact de lamelles de verre sont infestées avec des spores de *E. cuniculi*. Une heure après l'infestation, les spores extracellulaires résiduelles sont éliminées. Après une durée totale d'incubation de 24, 48, ou 72 h, les cellules sont lavées au PBS et fixées par le méthanol froid (20 min à -80°C). Dans le cas de spores extracellulaires, des frottis sont réalisés sur des lames de verre recouvertes de polylysine puis fixés comme précédemment.

Après une perméabilisation de 30 min en présence de PBS-Triton X 100 0,5 %, une étape de saturation en PBS-lait écrémé 5 % pendant 1 heure est réalisée. Le matériel biologique est ensuite incubé pendant 1 h avec les anticorps primaires dilués au 1/100 en PBS-Triton X-100 0,1 %. Après plusieurs rinçages au PBS-Triton X-100 0,1 %, les cellules sont incubées 1 h avec les anticorps secondaires (anti-IgG de souris ou de lapin) conjugués au fluorochrome Alexa 488 (dilué au 1/1000, Molecular Probes) ou TRITC (dilué au 1/400, Sigma). Afin de mieux visualiser les foyers infectieux, les noyaux des spores sont révélés au DAPI. Les préparations sont observées à l'aide d'un microscope à épifluorescence LEICA DMR.

II- Résultats

II- 1 Réactivité des anticorps anti-ECU01_0110

La séquence Y111_ENCCU, correspondant à une protéine InterB putative codée par le CDS ECU01_0110, a été exprimée après exclusion de son signal peptide chez *E.*

coli. L'antisérum anti-ECU01_0110 produit contre la protéine recombinante réagit avec de nombreuses bandes du profil protéique de l'extrait sporal d'*E. cuniculi*, transféré sur membrane PVDF (**Figure 43**). Deux bandes sont particulièrement réactives dans la zone 70-80 kDa. La taille de la bande inférieure est proche de celle calculée pour la protéine Y111_ENCCU (68,5 kDa). Il est donc possible qu'au moins une protéine InterB soit exprimée. Les expériences d'immunolocalisation avec le même antisérum n'ont montré qu'un marquage très diffus de la spore, apparemment peu spécifique.

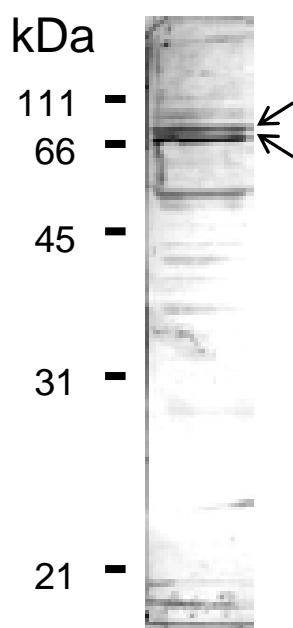


Figure 43: Western blot de protéines sporales d'*E. cuniculi* après traitement par des anticorps polyclonaux anti-ECU01_0110. Deux bandes (flèches) sont très réactives.

II-2 Expression de différentes régions de newECU03 chez *E. coli*

Les amplifications PCR pour les quatre régions sélectionnées dans la séquence *newECU03* ont fourni des produits ayant les tailles espérées (**Figure 44**).

Les différents fragments amplifiés ont été clonés dans le vecteur pGEX-4T1-His puis les plasmides recombinants ont servi à la transformation de bactéries BL21+. Après induction par l'IPTG de l'expression, les protéines ont été analysées en SDS-PAGE. L'expression de la protéine GST seule a aussi été réalisée, à titre de témoin.

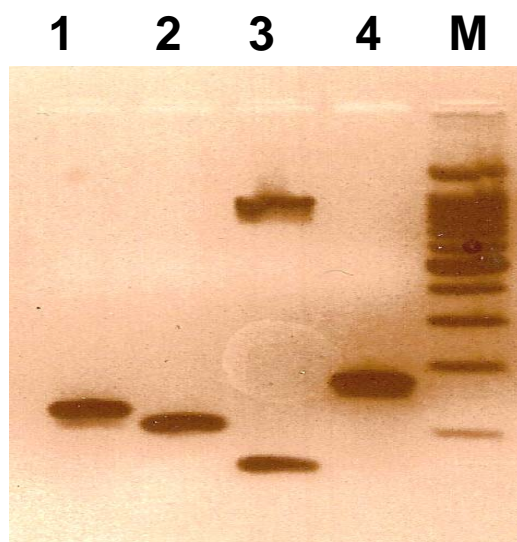


Figure 44 : Profils en gel d'agarose des produits d'amplification PCR pour les quatre fragments d'intérêt du gène *newECU03*. **M** : marqueurs de taille moléculaire (Promega).

1/ fragment *newECU03_PR1*, amplifié par le couple *5newEcu03spec1* Vs *3newEcu03spec2*.

2 / fragment *newECU03_PR2*, amplifié par *5newEcu03spec1* Vs *3newEcu03spec3*.

3/ fragment *newECU03_PR3*, amplifié par *5newEcu03spec1* Vs *3newEcu03gener1* (936 pb, bande supérieure).

4/ fragment *newECU03_PR4*, amplifié par *5newEcu03gener2* Vs *3newEcu03gener2*.

Seuls les fragments correspondant aux amplicons 1 (*newECU03_PR1*) et 4 (*newECU03_PR4*) ont pu être exprimés chez *Escherichia coli*. Les profils SDS-PAGE montrent que les protéines de fusion migrent à 31 kDa pour *newECU03_PR1/GST* et à 33 kDa pour *newECU03_PR4/GST* (**Figure 45**). Les bandes contenant ces protéines recombinantes ont été découpées puis purifiées et injectées à des souris. Les premiers antisérums collectés sont en cours d'utilisation pour des essais d'immunolocalisation.

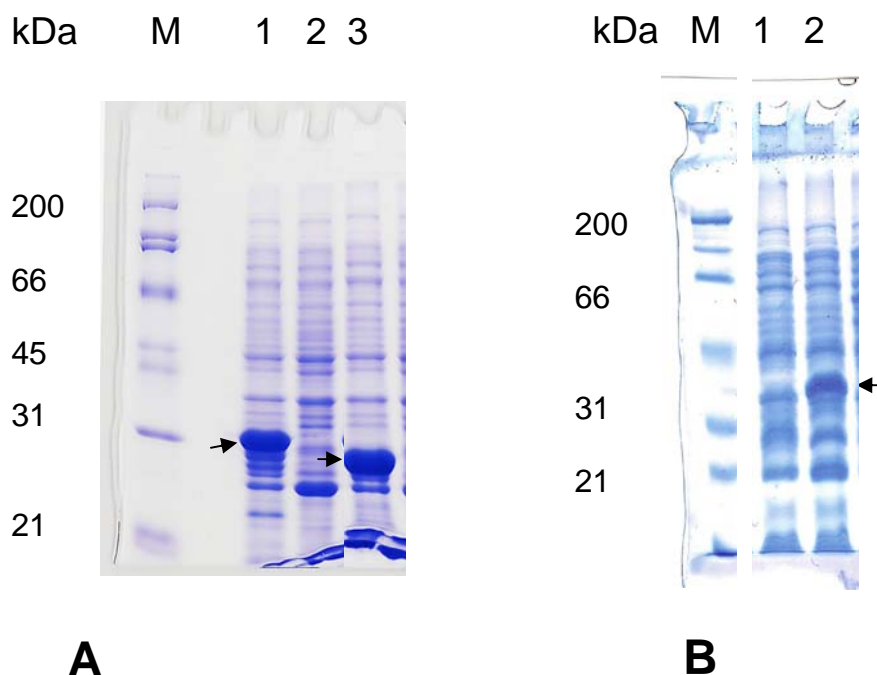


Figure 45: Expression chez *Escherichia coli* des protéines recombinantes correspondant aux amplicons 1 (A) et 4 (B). Les protéines de bactéries *E. coli* BL21+ transformées avec les vecteurs pGEX-4T1 recombinants sont analysées en SDS-PAGE 10%, après solubilisation en tampon Laemmli. M : marqueurs de taille moléculaire (Biorad).

A/ La piste 1 montre un profil avec la protéine exprimée (flèche) après induction par l'IPTG. La taille de 31 kDa pour cette protéine correspond bien à celle prévue pour la fusion de la GST (27 kDa) avec le peptide newECU03_PR1 (37 acides aminés, soit environ 4 kDa). Les pistes 2 et 3 représentent les profils témoins : 2/ sans induction par IPTG, 3/ expression de GST seule (flèche).

B/ La piste 1 correspond au témoin non induit. La piste 2 révèle l'expression d'une protéine proche de 33 kDa (flèche) représentative de la GST (27 kDa) fusionnée au peptide newECU03_PR4 (52 acides aminés, soit 5,7 kDa).

Discussion et perspectives

Discussion et perspectives

Dans un contexte de parasitologie, les protéines codées par des gènes en position subterminale sur les chromosomes sont d'importance centrale, celles-ci étant souvent associées à la pathogénie des espèces parasites, quel que soit le phylum auquel elles appartiennent (Barry et al, 2003). Les modalités de l'expression des gènes *var* chez l'apicomplexe *Plasmodium falciparum* (agent du paludisme) et des gènes *vsg* chez le kinétoplastidé *Trypanosoma brucei* (maladie du sommeil) sont sans aucun doute les mieux caractérisées. Les travaux sur ces familles multigéniques étaient dictés par le souci de justifier la grande variabilité d'une protéine de surface immunodominante (PfEMP1 chez *P. falciparum*, VSG chez *T. brucei*) et de déterminer le mécanisme assurant la commutation du variant exprimé.

Bien que n'ayant pu s'appuyer sur la connaissance d'un phénomène similaire de variation antigénique chez les microsporidies, la présente étude constitue une première contribution à l'analyse post-génomique de séquences répétées à localisation subterminale chez *Encephalitozoon cuniculi*. Nous avons démontré que l'une des familles multigéniques potentielles (famille *interB*) est susceptible d'être exprimée et se trouve représentée chez quelques autres microsporidies capables d'infecter l'Homme. Par rapport à ces résultats, nous allons orienter notre discussion autour de quatre points principaux :

- ✓ le contexte chromosomique et génétique du répertoire *interB*.
- ✓ la transcription des gènes *interB* et le mode de régulation.
- ✓ la localisation potentielle des protéines InterB.
- ✓ la relation avec la spécificité d'hôte.

I- Le répertoire *interB* chez *Encephalitozoon cuniculi*

Pour un eucaryote, *Encephalitozoon cuniculi* possède un génome très petit (2,9 Mb, Katinka et al, 2001). La présence de familles multigéniques chez ce parasite intracellulaire qui, à l'image d'autres microsporidies, s'est engagé dans une voie de réduction de son génome (Keeling et Slamovits, 2004), est assez surprenante. Quatre familles de gènes, *interAE*, *interB*, *interD* et *interC* ont été distinguées et sont principalement associées aux régions subterminales des chromosomes.

Ces familles apparaissent candidates à des gènes dits de « contingence » (Barry et al, 2003), gènes soumis à de fréquentes mutations et conférant des phénotypes alternatifs liés à des changements environnementaux. Chez les parasites eucaryotes, les gènes de contingence jouent un rôle majeur dans la protection contre les réponses immunes de l'organisme hôte, à travers la variabilité des protéines de surface exprimées. Dans le cas des microsporidies, la forme cellulaire assurant l'infestation est une spore couverte par une épaisse paroi riche en chitine mais le composant protéique majeur (SWP1) de la couche pariétale externe (exospore) chez *E. cuniculi* n'est codé que par un seul gène (Bohne et al, 2000). Le succès de la prolifération d'un parasite microsporidien exige cependant des interactions variées au cours d'une série de processus coordonnés: reconnaissance d'une cellule-cible adéquate, extrusion du tube polaire, internalisation du sporoplasme, protection contre la dégradation par les lysosomes, importation de nutriments et divers précurseurs pour les biosynthèses... Les modalités de la reconnaissance et de la pénétration du parasite sont encore débattues au niveau cytotologique. Les données moléculaires sont principalement limitées à l'identification de trois protéines de tube polaire codées par des gènes uniques et la mise en évidence de la prépondérance de courtes chaînes O-mannosylées dans les glycoprotéines (Taupin et al, 2006). Une implication de protéines issues de l'expression de familles multigéniques dans une phase particulière de l'invasion microsporidienne reste donc envisageable.

La famille *interB* qui a fait l'objet de cette étude est un répertoire composé de 36 gènes qu'on a pu subdiviser en 6 sous-familles sur la base d'homologie de séquences des protéines. Ces gènes ne sont absents que sur trois extrémités chromosomiques. Certains sont des pseudogènes (11 au total) codant pour des protéines tronquées. Les gènes *interB* codent pour une protéine d'environ 600 acides aminés et partagent la même orientation de transcription avec l'unité ADN_r (en direction des télomères). Contrairement aux gènes *var* (revue Frank et al, 2006), aucun intron n'a été détecté.

II- Transcription des gènes *interB* : quel contrôle ?

Nous avons utilisé l'approche RT-PCR pour étudier la transcription des gènes *interB*. Ce choix a été fait après une validation de la spécificité de nos différentes amorces (amorces spécifiques pour une amplification par sous-famille de gènes et amorces que nous qualifions de génériques car permettant d'amplifier tous les gènes du répertoire). Ainsi avec ces différentes amorces, nous avons montré que tous les gènes *interB* sont transcrits. Cependant, le séquençage aléatoire de clones (clonage des produits RT-PCR par utilisation d'amorces génériques) a révélé la prédominance du transcrit issu du gène que nous avons appelé

NewECU03 (localisé à l'extrémité b du chromosome 3). Ce résultat pose donc le problème de la régulation de la transcription de cette famille *interB*. Trois points méritent d'être discutés :

- une régulation de type épigénétique.
- la temporalité de l'expression de ces gènes
- le type de promoteur

L'hypothèse d'une régulation épigénétique est à considérer car les gènes *interB* occupent une position privilégiée sur les chromosomes, à savoir la région subterminale, juste en amont de l'unité ADN_r et représentant vraisemblablement une zone de transition hétérochromatine-euchromatine. On peut logiquement supposer une influence de la compaction chromatinienne sur l'expression de ces gènes. La prédiction de gènes pour des protéines de type Sir2 (*ECU03_0460*) chez *E. cuniculi* renforce cette idée. La protéine Sir2 a été particulièrement étudiée chez *S. cerevisiae*. Chez cette levure, Sir2 est associée à la répression transcriptionnelle de plusieurs locus et sa surexpression entraîne une désacétylation généralisée des histones. Cette enzyme ne présente d'activité histone désacétylase qu'en présence d'un co-facteur, le nicotinamide adénine dinucléotide (NAD). La désacétylation des histones par Sir2 est couplée à l'hydrolyse du NAD qui libère du nicotinamide et un autre métabolite, l'acétyl-ADP-ribose (Moazed et al, 2001). Par son activité histone désacétylase, Sir2 participe à l'hétérochromatisation. En effet, lors de ce processus, la libération des groupements acétyl, est suivie d'une méthylation des histones (par des méthylases comme Suv39H1). Cette modification est reconnue par le répresseur transcriptionnel HP1 (Heterochromatin Protein 1) chez la drosophile et les mammifères et entraîne la formation d'hétérochromatine (**Figure 46**) (Bannister et al, 2001; Lachner et al, 2001). Ces protéines sont dépourvues de domaine de liaison à l'ADN. Leur association avec l'hétérochromatine dépend de leur domaine amino-terminal (*chromo*) qui se fixe de façon très stable à l'histone H3 méthylée sur le groupement ϵ -aminé de la lysine 9 (Lys 9) (Nielsen et al, 2002).

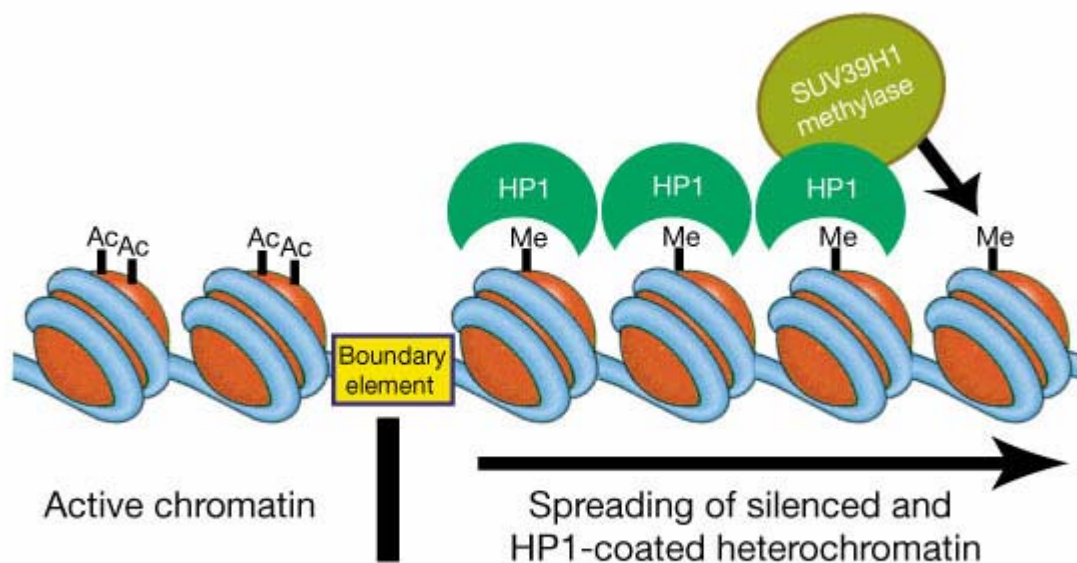


Figure 46: Processus d'hétérochromatisation de la chromatine sous l'influence de HP1.

Les histones acétylés marque une forme relaxée et donc active de la chromatine. La désacétylation de ces histones par Sir2 favorise une cascade d'événements, dont la méthylation d'histones (méthylase impliquée : SUV39H) permettant la fixation des protéines HP1, composants essentiels de l'hétérochromatine (selon **Bannister et al, 2001**).

Sachant qu'*in vitro*, l'activité de Sir2 dépend directement de la concentration de son co-facteur NAD, les gènes *interB* ont plus de chances d'être exprimés à un stade tardif du cycle (à partir du sporonte) avec une biosynthèse élevée (notamment la biosynthèse des phospholipides) qui va mobiliser une grande partie du NAD cellulaire au détriment de l'activité sir2.

La seconde question qui se pose est de savoir s'il existe ou non une relation entre l'expression d'*interB* et le stade de développement. En travaillant sur de l'ARN total extrait de cellules infestées sur plusieurs semaines, les ARN de tous les stades parasitaires sont récupérés. Cependant, le stade prédominant est la spore mature. Cela pourrait expliquer le fait qu'on ait obtenu, au cours de trois expériences différentes, le même transcrit majoritaire. Une étude de la transcription chez différents stades de développement du parasite par une méthode d'hybridation *in situ* avec des sondes *interB* nous renseignera plus clairement sur la temporalité de l'expression de ces gènes. Il faudra également étudier le répertoire transcrit par une approche RT-PCR quantitative afin de vérifier les niveaux de transcription des différents

gènes. L'étude du répertoire transcrit sur les autres microsporidies positives aux gènes *interB* est à entreprendre, surtout pour les modèles *Brachiola sp* et *Vittaforma corneae*.

Pour ce qui concerne la recherche de promoteurs, soulignons qu'aucune boîte TATA ou CAAT n'a été identifiée en amont de chaque gène *interB*, ce qui pose la question du type d'ARN polymérase assurant la transcription. Les gènes *interB* sont situés juste en aval de l'ADNr, avec une même orientation (transcription orientée vers le télomère). L'ARN polymérase I serait-elle alors impliquée? La transcription polycistronique par l'ARN polymérase I est bien réalisée chez *Trypanosoma brucei*, pour l'expression des gènes *vsg* (Günzl et al, 2003). Il serait donc intéressant d'étudier la structure des transcrits *interB* par « une marche sur le transcrit » en utilisant des amorces allant toujours plus loin. On pourra utiliser l'approche Race PCR (techniques Race 3' et 5') pour les extrémités.

III- Protéines InterB et localisation cellulaire

Les séquences protéiques InterB ne présentent pas d'homologies significatives avec des protéines ou des domaines fonctionnels connus. De la comparaison de ces séquences, nous avons déduit une organisation de base comprenant une région N-terminale avec au moins un domaine transmembranaire potentiel, une région centrale hypervariable et une région C-terminale conservée. Un peptide signal N-terminal est parfois prédit avec un score élevé. L'hypothèse d'une protéine adressée à une membrane est envisageable mais nous ne disposons actuellement d'aucune preuve expérimentale.

L'utilisation d'anticorps polyclonaux afin d'étudier la localisation au niveau cellulaire des protéines InterB a donné des signaux très faibles, non spécifiques (résultats non montrés). Diverses raisons peuvent être invoquées pour justifier ces résultats: entre autres, une perméabilisation difficile des stades sporogoniques revêtus d'une paroi plus ou moins épaisse, épitopes séquentiels non démasqués, faible spécificité de l'antisérum, faible affinité des anticorps produits pour les protéines InterB, l'immunogénicité de ces protéines. Par rapport à leur possible localisation, nous émettons deux hypothèses:

- une localisation au niveau de la membrane plasmique du parasite.
- un adressage à la cellule-hôte.

Dans cette seconde hypothèse, la sécrétion de protéines destinées à la cellule hôte par un parasite est un phénomène bien connu chez les parasites intracellulaires, particulièrement chez les apicomplexes (*P. falciparum*, *T. gondii*). En se référant au cas des protéines PfEMP-1 (Plasmodium falciparum Erythrocyte Membrane Protein- 1) de *P. falciparum*, il pourrait être imaginé une exportation en direction de la surface de la cellule-hôte. Cependant, il faut

souligner que *Plasmodium* et les microsporidies diffèrent très significativement par le type cellulaire infecté. L'hématie où pénètre *Plasmodium* est une cellule anucléée, sans organites cytoplasmiques et métaboliquement peu active, ce qui n'est pas le cas des cellules susceptibles d'être envahies par les microsporidies. La découverte récente chez un phytoparasite *Phytophthora infestans* (parasite de pomme de terre) de protéines adressées à la cellule-hôte pourvues de la même séquence d'adressage dite HT (host targeting) que celle des protéines de *P. falciparum* (Bhattacharjee et al, 2006) montre que ce mécanisme est indépendant de la nature de la cellule-hôte. Ainsi, il peut y avoir conservation durant l'évolution, de mécanismes d'interaction hôte-parasite à une échelle moléculaire pouvant être intégrés dans des voies variables. Ces signaux d'adressage ne sont pas présents chez *E. cuniculi*. Par contre, il n'est pas impossible que le trafic des protéines d'origine microsporidienne se fasse grâce à des transporteurs ABC (Cornillot et al, 2002). Certains de ces transporteurs ont des signaux d'adressage de type peptide signal et peuvent se retrouver aussi bien à la surface du parasite qu'au niveau de la membrane de la vacuole parasitophore, d'autant plus que la membrane des mérontes est étroitement liée à la membrane vacuolaire.

Quoi qu'il en soit, des études immunocytochimiques à l'échelle ultrastructurale sont à entreprendre. Il serait évidemment souhaitable de préparer des anticorps plus spécifiquement dirigés contre la protéine correspondant au transcrit majoritaire et, peut-être mieux, contre les motifs les plus caractéristiques de la famille InterB.

III- Gènes *interB*: relation avec la spécificité d'hôte ?

La présence du répertoire *interB* chez des microsporidies parasites de moustique (genre *Brachiola*) et de l'Homme (genres *Encephalitozoon* et *Vittaforma*) nous permet d'émettre des hypothèses sur l'éventuel rôle des produits de ces gènes dans les interactions hôte-parasite. Il est vrai que le répertoire révèle une voie d'évolution inattendue au regard des données phylogénétiques. Cependant, ces résultats ne sont pas en contradiction avec les données biologiques qui montrent que les microsporidies du genre *Brachiola* sont capables d'infester des mammifères dont l'Homme (études expérimentales de Trammer et al, 1997; atteintes de la cornée observées chez des personnes immunocompétentes ou immunodéprimées, Visvesvara et al, 1999; Coyle et al, 2004). La transmission du moustique au mammifère a été décrite pour *Trachipleistophora spp*, un genre très proche du genre *Vavraia* (Cheney et al, 2001) qui reste parasite de moustique. Il serait intéressant de rechercher les gènes *interB* chez ces microsporidies afin de vérifier si la voie d'évolution suivie par *Brachiola* et *Encephalitozoon* concerne un groupe plus vaste.

Ceci pose la question du passage du parasite entre insectes et mammifères et du rôle potentiel des protéines InterB dans ce contexte. L'hypothèse d'une transmission due à l'acquisition du repas sanguin des moustiques chez des mammifères, a été proposée pour expliquer le franchissement de la barrière d'espèce pour *Brachiola* par exemple (revue Didier, 2005). Cependant, la récente description d'un hérisson de Madagascar parasité par une espèce du genre *Trachipleistophora* (Vavra et al, 2006) montre que la transmission par voie orale est possible.

Le passage à un hôte mammifère pourrait être lié à la simplification du cycle biologique. En effet, contrairement à celles du genre *Amblyospora*, les espèces des genres *Brachiola*, *Trachipleistophora* et *Vavraia* sont incapables de méiose et leur transmission se fait directement par l'environnement sans passer par un hôte intermédiaire. A ce premier niveau d'analyse, il est donc improbable que les protéines InterB soient impliquées dans une phase sexuée du cycle. Compte tenu des différentes voies d'entrée possibles du parasite, un rôle dans les premières étapes de l'infestation de l'hôte semble aussi exclu.

Certaines microsporidies parasitant l'Homme sont phylogénétiquement proches de celles parasitant des poissons. Des cas d'infestation de personnes immunodéprimées par des membres du genre *Pleistophora sp* (Chup et al, 1993) ou *Enterocytozoon sp* (Orenstein JM, 1991) qui est particulièrement proche de *Nucleospora salmonis* infectant des saumons. Ceci va dans le sens de l'hypothèse d'un passage de la barrière d'espèce lié à l'alimentation. Nous n'avons pu rechercher les gènes *interB* chez ces parasites, mais au vu des données phylogénétiques et épidémiologiques, leur présence est peu probable.

Dans tous les cas une recherche des gènes *interB* dans les différents genres parasitant l'Homme est à entreprendre afin d'établir un lien plus clair entre ces gènes et la spécificité d'hôte.

Références bibliographiques

Références bibliographiques

- Abrahamsen MS, Templeton TJ, et al. 2004.** Complete genome sequence of the apicomplexan, *Cryptosporidium parvum*. *Science*. 304(5669):441-5.
- Adam RD. 1991.** The biology of *Giardia spp.* *Microbiol Rev*. 55(4):706-32.
- Adam RD. 2000.** The *Giardia lamblia* genome. *Int J Parasitol*. 30(4):475-84.
- Adams JH, Fang X, Kaslow DC, Miller LH. 1992.** Identification of a cryptic intron in the *Plasmodium vivax* Duffy binding protein gene. *Mol Biochem Parasitol*. ;56(1):181-3.
- Ahmad K, Henikoff S. 2002.** Histone H3 variants specify modes of chromatin assembly. *Proc Natl Acad Sci U S A*. 99 Suppl 4:16477-84.
- Ajioka JW, Fitzpatrick JM, Reitter CP. 2001.** *Toxoplasma gondii* genomics: shedding light on pathogenesis and chemotherapy. *Expert Rev Mol Med*. 2001:1-19.
- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG. 1981.** Sequence and organization of the human mitochondrial genome. *Nature* 290(5806):457-65.
- Andreadis TG. 2005.** Evolutionary strategies and adaptations for survival between mosquito-parasitic microsporidia and their intermediate copepod hosts: a comparative examination of *Amblyospora connecticus* and *Hyalinocysta chapmani* (Microsporidia: Amblyosporidae). *Folia Parasitol (Praha)*.;52(1-2):23-35.
- Arkhipova IR, Morrison HG. 2001.** Three retrotransposon families in the genome of *Giardia lamblia*: two telomeric, one dead. *Proc Natl Acad Sci U S A*. 98(25):14497-502.
- Bacchi CJ and Weiss LM 2004.** Chemotherapy of microsporidiosis: benzimidazoles, fumagillin and polyamine analogues. In *Opportunistic infections: Toxoplasma, Sarcocystis, and Microsporidia* (Eds Lindsay DS and Weiss LM), *World Class Parasites*, vol. 9, pp 159-188.
- Bachand F, Triki I, Autexier C. 2001.** Human telomerase RNA-protein interactions. *Nucleic Acids Res*. 29(16):3385-93.
- Bhattacharjee S, Hiller NL, Liolios K, Win J, Kanneganti TD, Young C, Kamoun S, Haldar K. 2006.** The malarial host-targeting signal is conserved in the Irish potato famine pathogen. *PLoS Pathog*. 2(5):e50. Epub 2006 May 26.
- Bannister AJ, Zegerman P, Partridge JF, Miska EA, Thomas JO, Allshire RC, Kouzarides T. 2001.** Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature*. 410(6824):120-4.
- Barry JD, Ginger ML, Burton P, McCulloch R. 2003.** Why are parasite contingency genes often associated with telomeres? *Int J Parasitol*. 33(1):29-45.
- Baruch DI. 1999.** Adhesive receptors on malaria-parasitized red cells. *Baillieres Best Pract Res Clin Haematol*. ;12(4):747-61.

- Bastien P, Blaineau C, Pages M. 1992.** *Leishmania*: sex, lies and karyotype.
Parasitol Today. 8(5):174-7.
- Baum M, Ngan VK, Clarke L. 1994.** The centromeric K-type repeat and the central core are together sufficient to establish a functional *Schizosaccharomyces pombe* centromere.
Mol Biol Cell. 5(7):747-61.
- Baumann P, Cech TR. 2001.** Pot1, the putative telomere end-binding protein in fission yeast and humans.
Science. 292(5519):1171-5.
- Baxevanis AD, Landsman D. 1996.** Histone Sequence Database: a compilation of highly-conserved nucleoprotein sequences.
Nucleic Acids Res 24(1):245-7.
- Beattie TL, Zhou W, Robinson MO, Harrington L. 2001.** Functional multimerization of the human telomerase reverse transcriptase.
Mol Cell Biol. 21(18):6151-60.
- Bechert T, Heck S, Fleig U, Diekmann S, Hegemann JH. 1999.** All 16 centromere DNAs from *Saccharomyces cerevisiae* show DNA curvature.
Nucleic Acids Res. 27(6):1444-9.
- Becker M, Aitcheson N, Byles E, Wickstead B, Louis E, Rudenko G. 2004.** Isolation of the repertoire of VSG expression site containing telomeres of *Trypanosoma brucei* 427 using transformation-associated recombination in yeast.
Genome Res.;14(11):2319-29.
- Becnel JJ, White SE, Shapiro AM. 2005.** Review of microsporidia-mosquito relationships: from the simple to the complex.
Folia Parasitol (Praha);52(1-2):41-50.
- Beja O, Schwartz D, Michaeli S. 1994.** Karyotype analysis of the monogenetic trypanosomatid *Leptomonas collosoma*.
Mol Biochem Parasitol. 66(1):71-81.
- Bendich AJ, Drlica K. 2000.** Prokaryotic and eukaryotic chromosomes: what's the difference?
Bioessays. 22(5):481-6.
- Berendt AR, Ferguson DJ, Newbold CI. 1990.** Sequestration in *Plasmodium falciparum* malaria: sticky cells and sticky problems.
Parasitol Today.;6(8):247-54.
- Berriman M, Ghedin E, Hertz-Fowler C, et al. 2005.** The genome of the African trypanosome *Trypanosoma brucei*.
Science. ;309(5733):416-22.
- Biderre C, Metenier G, Vivares CP. 1999.** Sequencing of several protein-coding genes of the chromosome X from the microsporidian *Encephalitozoon cuniculi*.
J Eukaryot Microbiol. 46(5):27S-28S.
- Biderre, C., Mathis, A., Deplazes, P., Weber, R., Méténier, G. & Vivarès, C. P. (1999b).** Molecular karyotype diversity in the microsporidian *Encephalitozoon cuniculi*.
Parasitology **118**, 439-45
- Biderre, C., Pagès, M., Méténier, G., Canning, E. U. & Vivarès, C. P. (1995).** Evidence for the smallest nuclear genome (2.9 Mb) in the microsporidium *Encephalitozoon cuniculi*.
Mol Biochem Parasitol **74**, 229-31.
- Biderre, C., Pagès, M., Méténier, G., David, D., Bata, J., Prensier, G. & Vivarès, C. P. (1994).** On small genomes in eukaryotic organisms: molecular karyotypes of two microsporidian species (Protozoa) parasites of vertebrates.

- C R Acad Sci III* **317**, 399-404.
- Bienz M, Siles-Lucas M, Wittwer P, Muller N. 2001.** vsp gene expression by *Giardia lamblia* clone GS/M-83-H7 during antigenic variation in vivo and in vitro.
Infect Immun. 69(9):5278-85.
- Bigliardi E, Sacchi L. 2001.** Cell biology and invasion of the microsporidia.
Microbes Infect.;3(5):373-9.
- Blackburn EH, Gall JG. 1978.** A tandemly repeated sequence at the termini of the extrachromosomal ribosomal RNA genes in *Tetrahymena*.
J Mol Biol. 120: 33-53.
- Blackburn EH. 2001.** Switching and signaling at the telomere.
Cell. 106(6):661-73.
- Blattner FR, Plunkett G 3rd, Bloch CA, et al. 1997.** The complete genome sequence of *Escherichia coli* K-12.
Science.;277(5331):1453-74.
- Blum ML, Down JA, Gurnett AM, Carrington M, Turner MJ, Wiley DC. 1993.** A structural motif in the variant surface glycoproteins of *Trypanosoma brucei*.
Nature. ;362(6421):603-9.
- Bohne W, Ferguson DJ, Kohler K, Gross U. 2000.** Developmental expression of a tandemly repeated, glycine- and serine-rich spore wall protein in the microsporidian pathogen *Encephalitozoon cuniculi*.
Infect Immun. 68(4):2268-75.
- Borst P, Bitter W, Blundell PA, Chaves I, Cross M, Gerrits H, van Leeuwen F, McCulloch R, Taylor M, Rudenko G. 1998.** Control of VSG gene expression sites in *Trypanosoma brucei*.
Mol Biochem Parasitol. ;91(1):67-76.
- Borst P. 2002.** Antigenic variation and allelic exclusion.
Cell. ;109(1):5-8.
- Botstein D, White RL, Skolnick M, Davis RW. 1980.** Construction of a genetic linkage map in man using restriction fragment length polymorphisms.
Am J Hum Genet. 32(3):314-31.
- Bowman S, Lawson D, Basham D, et al. 1999.** The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*.
*Nature.*400(6744):532-8.
- Bringaud F, Biteau N, Melville SE, Hez S, El-Sayed NM, Leech V, Berriman M, Hall N, Donelson JE, Baltz T. 2002.** A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of *Trypanosoma brucei*.
Eukaryot Cell. 1(1):137-51.
- Britten RJ. 1998.** Precise sequence complementarity between yeast chromosome ends and two classes of just-subtelomeric sequences.
Proc Natl Acad Sci U S A. 95(11):5906-12.
- Brugere JF, Cornillot E, Bourbon T, Metenier G, Vivares CP. 2001.** Inter-strain variability of insertion/deletion events in the *Encephalitozoon cuniculi* genome: a comparative KARD-PFGE analysis.
J Eukaryot Microbiol.; Suppl:50S-55S.
- Brugere JF, Cornillot E, Metenier G, Bensimon A, Vivares CP (2000a).** *Encephalitozoon cuniculi* (Microspora) genome: physical map and evidence for telomere-associated rDNA units on all chromosomes.
Nucleic Acids Res **28**: 2026-2033

- Brugere JF, Cornillot E, Metenier G, Vivares CP (2000b).** In-gel DNA radiolabelling and two-dimensional pulsed field gel electrophoresis procedures suitable for fingerprinting and mapping small eukaryotic genomes.
Nucleic Acids Res **28**: E48
- Brugere JF, Cornillot E, Metenier G, Vivares CP (2000c).** Occurrence of subtelomeric rearrangements in the genome of the microsporidian parasite *Encephalitozoon cuniculi*, as revealed by a new fingerprinting procedure based on two-dimensional pulsed field gel electrophoresis.
Electrophoresis **21**: 2576-2581
- Burke DT. 1991.** The role of yeast artificial chromosome clones in generating genome maps.
Curr Opin Genet Dev. 1(1):69-74.
- Burri L, Williams BA, Bursac D, Lithgow T, Keeling PJ. 2006.** Microsporidian mitosomes retain elements of the general mitochondrial targeting system.
Proc Natl Acad Sci U S A.;103(43):15916-20. Epub 2006 Oct 16.
- Bustamante C, Gurrieri S, Smith SB. 1993.** Towards a molecular description of pulsed-field gel electrophoresis.
Trends Biotechnol. 11(1):23-30.
- Bustin M. 1999.** Regulation of DNA-dependent activities by the functional motifs of the high-mobility-group chromosomal proteins.
Mol Cell Biol. 19(8):5237-46.
- Calderwood MS, Gannoun-Zaki L, Wellems TE, Deitsch KW. 2003.** *Plasmodium falciparum* var genes are regulated by two regions with separate promoters, one upstream of the coding region and a second within the intron.
J Biol Chem. ;278(36):34125-32.
- Cali A. and Takvorian P. 1999.** Developmental morphology and life cycles of the microsporidia. In: The Microsporidia and Microsporidiosis, edited by M. Wittner and L. Weiss (American Society for microbiology, Washington, DC), P. 85-128.
- Carle GF, Frank M, Olson MV. 1986.** Electrophoretic separations of large DNA molecules by periodic inversion of the electric field.
Science. 232(4746):65-8.
- Carranza PG, Feltes G, Ropolo A, Quintana SM, Touz MC, Lujan HD. 2002.** Simultaneous expression of different variant-specific surface proteins in single *Giardia lamblia* trophozoites during encystation.
Infect Immun. 70(9):5265-8.
- Cathcart R. 1990.** Advances in automated DNA sequencing.
Nature. 347(6290):310.
- Cavalier-Smith T. 1998.** A revised six-kingdom system of life.
Biol Rev Camb Philos Soc. 73(3):203-66.
- Chen Q, Fernandez V, Sundstrom A, Schlichtherle M, Datta S, Hagblom P, Wahlgren M. 1998.** Developmental selection of var gene expression in *Plasmodium falciparum*.
Nature. 394(6691):392-5.
- Cheney SA, Lafranchi-Tristem NJ, Canning EU. 2001.** Serological differentiation of microsporidia with special reference to *Trachipleistophora hominis*.
Parasite. 8(2):91-7.
- Chu G, Vollrath D, Davis RW. 1986.** Separation of large DNA molecules by contour-clamped homogeneous electric fields.
Science. 234(4783):1582-5.

- Chup G L, Alroy J, Adelman LS, et al. 1993.** Myositis due to *Pleistophora* (Microsporidia) in a patient with AIDS.
Clin Infect Dis 16:15-21.
- Clarke L, Baum MP. 1990.** Functional analysis of a centromere from fission yeast: a role for centromere-specific repeated DNA sequences.
Mol Cell Biol. 10(5):1863-72.
- Clarke L, Carbon J. 1980.** Isolation of the centromere-linked CDC10 gene by complementation in yeast.
Proc Natl Acad Sci U S A. 77(4):2173-7.
- Clarke L. 1990.** Centromeres of budding and fission yeasts.
Trends Genet. 6(5):150-4.
- Cohn M, McEachern MJ, Blackburn EH. 1998.** Telomeric sequence diversity within the genus *Saccharomyces*.
Curr Genet. 33(2):83-91.
- Conway C, McCulloch R, Ginger ML, Robinson NP, Browitt A, Barry JD. 2002.**
Ku is important for telomere maintenance, but not for differential expression of telomeric VSG genes, in African trypanosomes.
J Biol Chem. 277(24):21269-77.
- Cooper JP, Watanabe Y, Nurse P. 1998.** Fission yeast Taz1 protein is required for meiotic telomere clustering and recombination.
Nature. 392(6678):828-31.
- Cormack BP, Ghori N, Falkow S. 1999.** An adhesin of the yeast pathogen *Candida glabrata* mediating adherence to human epithelial cells.
Science. 285(5427):578-82.
- Cornelissen AW, Overdulve JP, van der Ploeg M. 1984.** Determination of nuclear DNA of five eucoccidian parasites, *Isospora* (*Toxoplasma*) *gondii*, *Sarcocystis cruzi*, *Eimeria tenella*, *E. acervulina* and *Plasmodium berghei*, with special reference to gamontogenesis and meiosis in *I.(T.) gondii*.
Parasitology. 88 (Pt 3):531-53.
- Cornillot E, Keller B, Cushion MT, Metenier G, Vivares CP. 2002.** Fine analysis of the *Pneumocystis carinii* f. sp. *carinii* genome by two-dimensional pulsed-field gel electrophoresis.
Gene. 293(1-2):87-95.
- Cornillot E, Metenier G, Vivares CP, Dassa E. 2002.** Comparative analysis of sequences encoding ABC systems in the genome of the microsporidian *Encephalitozoon cuniculi*.
FEMS Microbiol Lett. :210(1):39-47.
- Coyle CM, Weiss LM, Rhodes LV 3rd, Cali A, Takvorian PM, Brown DF, Visvesvara GS, Xiao L, Naktin J, Young E, Gareca M, Colasante G, Wittner M. 2004.** Fatal myositis due to the microsporidian *Brachiola algerae*, a mosquito pathogen.
N Engl J Med.;351(1):42-7.
- Curgy, J., Vavra, J. & Vivarès, C. (1980).** Presence of ribosomal RNAs with prokaryotic properties in Microsporidia, eukaryotic organisms.
Biol Cell 38, 49-52.
- Dai Q, Restrepo BI, Porcella SF, Raffel SJ, Schwan TG, Barbour AG. 2006.** Antigenic variation by *Borrelia hermsii* occurs through recombination between extragenic repetitive elements on linear plasmids.
Mol Microbiol. 60(6):1329-43.
- Daniels DL, Blattner FR. 1987.** Mapping using gene encyclopaedias.
Nature. 325(6107):831-2.

- de Lange T. 2001.** Cell biology. Telomere capping--one strand fits all.
Science. 292(5519):1075-6.
- de Lange T. 2002.** Protection of mammalian telomeres.
Oncogene. 21(4):532-40.
- Deitsch KW, Calderwood MS, Wellems TE. 2001.** Malaria. Cooperative silencing elements in *var* genes.
Nature. 412(6850):875-6.
- Deitsch KW, del Pinal A, Wellems TE. 1999.** Intra-cluster recombination and *var* transcription switches in the antigenic variation of *Plasmodium falciparum*.
Mol Biochem Parasitol.;101(1-2):107-16.
- Delarbre S, Gatti S, Scaglia M, Drancour M. 2001.** Genetic diversity in the microsporidian *Encephalitozoon hellem* demonstrated by pulsed-field gel electrophoresis.
J Eukaryot Microbiol.;48(4):471-4.
- Deshpande AM, Newlon CS. 1992.** The ARS consensus sequence is required for chromosomal origin function in *Saccharomyces cerevisiae*.
*Mol Cell Biol.*12(10):4305-13.
- Didier ES. 2005.** Microsporidiosis: an emerging and opportunistic infection in humans and animals.
Acta Trop.;94(1):61-76.
- Didier, E. S., Vossbrinck, C. R., Baker, M. D., Rogers, L. B., Bertucci, D. C. & Shadduck, J. A. (1995).** Identification and characterization of three *Encephalitozoon cuniculi* strains.
Parasitology **111**, 411-421.
- Diffley JF, Cocker JH. 1992.** Protein-DNA interactions at a yeast replication origin.
Nature. 357(6374):169-72.
- Diffley JF, Stillman B. 1990.** The initiation of chromosomal DNA replication in eukaryotes.
Trends Genet. 6(12):427-32.
- Diffley JF. 1996.** Once and only once upon a time: specifying and regulating origins of DNA replication in eukaryotic cells.
Genes Dev. 10(22):2819-30.
- Dimitri P, Corradini N, Rossi F, Verni F. 2005.** The paradox of functional heterochromatin.
Bioessays 27(1):29-41.
- Douglas S, Zauner S, Fraunholz M, Beaton M, Penny S, Deng LT, Wu X, Reith M, Cavalier-Smith T, Maier UG. 2001.** The highly reduced genome of an enslaved algal nucleus.
Nature. ;410(6832):1091-6.
- Dujon B, Sherman D, Fischer G, et al. 2004.** Genome evolution in yeasts.
Nature. 430(6995):35-44.
- Duraisingh MT, Voss TS, Marty AJ, et al 2005.** Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*.
Cell. ;121(1):13-24.
- Dynan WS, Yoo S. 1998.** Interaction of Ku protein and DNA-dependent protein kinase catalytic subunit with nucleic acids.
Nucleic Acids Res. 26(7):1551-9.
- Edlind TD, Li J, Visvesvara GS, Vodkin MH, McLaughlin GL, Katiyar SK. 1996.** Phylogenetic analysis of beta-tubulin sequences from amitochondrial protozoa.
Mol Phylogenet Evol.;5(2):359-67.

- Eichinger L, Pachebat JA, Glockner G, et al. 2005.** The genome of the social amoeba *Dictyostelium discoideum*. *Nature*. 435(7038):43-57.
- El-Sayed NM, Hegde P, Quackenbush J, Melville SE, Donelson JE. 2000.** The African trypanosome genome. *Int J Parasitol*. 30(4):329-45.
- Ey PL, Khanna KK, Manning PA, Mayrhofer G. 1993.** A gene encoding a 69-kilodalton major surface protein of *Giardia intestinalis* trophozoites. *Mol Biochem Parasitol*. 58(2):247-57.
- Felsenfeld G, Boyes J, Chung J, Clark D, Studitsky V. 1996.** Chromatin structure and gene expression. *Proc Natl Acad Sci U S A*. ;93(18):9384-8.
- Feng J, Funk WD, Wang SS, Weinrich SL, Avilion AA, Chiu CP, Adams RR, Chang E, Allsopp RC, Yu J, et al. 1995.** The RNA component of human telomerase. *Science*. 269(5228):1236-41.
- Feuerbach F, Galy V, Trelles-Sticken E, Fromont-Racine M, Jacquier A, Gilson E, Olivo-Marin JC, Scherthan H, Nehrbass U. 2002.** Nuclear architecture and spatial positioning help establish transcriptional states of telomeres in yeast. *Nat Cell Biol*. 4(3):214-21.
- Firtel RA, Cockburn A, Frankel G, Hershfield V. 1976.** Structural organization of the genome of *Dictyostelium discoideum*: analysis by EcoRI restriction endonuclease. *J Mol Biol*. 102(4):831-52.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al. 1995.** Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269(5223):496-512.
- Florida G, Zatterale A, Zuffardi O, Tyler-Smith C. 2000.** Mapping of a human centromere onto the DNA by topoisomerase II cleavage. *EMBO Rep*. 1(6):489-93.
- Fournier P, Abbas A, Chasles M, Kudla B, Ogrydziak DM, Yaver D, Xuan JW, Peito A, Ribet AM, Feynerol C, et al. 1993.** Colocalization of centromeric and replicative functions on autonomously replicating sequences isolated from the yeast *Yarrowia lipolytica*. *Proc Natl Acad Sci U S A*. 90(11):4912-6.
- Frank M, Deitsch K. 2006.** Activation, silencing and mutually exclusive expression within the var gene family of *Plasmodium falciparum*. *Int J Parasitol*. ;36(9):975-85.
- Franzen C, Muller A. 1999.** Molecular techniques for detection, species differentiation, and phylogenetic analysis of microsporidia. *Clin Microbiol Rev*. ;12(2):243-85.
- Franzen C. 2005.** How do microsporidia invade cells? *Folia Parasitol (Praha)*. ;52(1-2):36-40.
- Fraser CM, Gocayne JD, White O, et al. 1995.** The minimal gene complement of *Mycoplasma genitalium*. *Science* 270(5235):397-403.

- Freitas-Junior LH, Bottius E, Pirrit LA, Deitsch KW, Scheidig C, Guinet F, Nehrbass U, Wellems TE, Scherf A. 2000.** Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*.;407(6807):1018-22.
- Freitas-Junior LH, Hernandez-Rivas R, et al. 2005.** Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell*.;121(1):25-36.
- Frixione, E., Ruiz, L., Cerbon, J. & Undeen, A. H. (1997).** Germination of *Nosema algerae* (Microspora) spores: conditional inhibition by D2O, ethanol and Hg²⁺ suggests dependence of water influx upon membrane hydration and specific transmembrane pathways. *J Eukaryot Microbiol* **44**, 109-16.
- Frixione, E., Ruiz, L., Santillan, M., de Vargas, L. V., Tejero, J. M. & Undeen, A. H. (1992).** Dynamics of polar filament discharge and sporoplasm expulsion by microsporidian spores. *Cell Motility and the Cytoskeleton* **22**, 38-50.
- Friz CT. 1968.** The biochemical composition of the free-living amoebae *Chaos chaos*, *Amoeba dubia* and *Amoeba proteus*. *Comp Biochem Physiol.* 26(1):81-90.
- Galy V, Olivo-Marin JC, Scherthan H, Doye V, Rascalou N, Nehrbass U. 2000.** Nuclear pore complexes in the organization of silent telomeric chromatin. *Nature*. 403(6765):108-12.
- Gannoun-Zaki L, Jost A, Mu J, Deitsch KW, Wellems TE. 2005.** A silenced *Plasmodium falciparum* var promoter can be activated in vivo through spontaneous deletion of a silencing element in the intron. *Eukaryot Cell*.;4(2):490-2.
- Gardner MJ, Bishop R, Shah T, et al. 2005.** Genome sequence of *Theileria parva*, a bovine pathogen that transforms lymphocytes. *Science*. 309(5731):134-7.
- Gardner MJ, Hall N, Fung E, et al. 2002.** Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 419(6906):498-511.
- Gatton ML, Peters JM, Fowler EV, Cheng Q. 2003.** Switching rates of *Plasmodium falciparum* var genes: faster than we thought? *Trends Parasitol.* ;19(5):202-8.
- Germot, A., Philippe, H. & Le Guyader, H. (1997).** Evidence for loss of mitochondria in Microsporidia from a mitochondrial- type HSP70 in *Nosema locustae*. *Mol Biochem Parasitol* **87**, 159-68.
- Gilbert DM. 2001.** Making sense of eukaryotic DNA replication origins. *Science*. 294(5540):96-100.
- Goffeau A, Barrell BG, Bussey Het al. 1996.** Life with 6000 genes. *Science* 275(5303):1051-2.
- Gottschling DE, Aparicio OM, Billington BL, Zakian VA. 1990.** Position effect at *S. cerevisiae* telomeres: reversible repression of Pol II transcription. *Cell*. 63(4):751-62.

- Gunzl A, Bruderer T, Laufer G, Schimanski B, Tu LC, Chung HM, Lee PT, Lee MG. 2003.** RNA polymerase I transcribes procyclin genes and variant surface glycoprotein gene expression sites in *Trypanosoma brucei*. *Eukaryot Cell.* ;2(3):542-51.
- Guo B, Styles CA, Feng Q, Fink GR. 2000.** A *Saccharomyces* gene family involved in invasive growth, cell-cell adhesion, and mating. *Proc Natl Acad Sci U S A.* ;97(22):12158-63.
- Gurrieri S, Smith SB, Bustamante C. 1999.** Trapping of megabase-sized DNA molecules during agarose gel electrophoresis. *Proc Natl Acad Sci U S A.* 96(2):453-8.
- Hardy CF, Sussel L, Shore D. 1992.** A RAP1-interacting protein involved in transcriptional silencing and telomere length regulation. *Genes Dev.* 6(5):801-14.
- Hazard, E. I. & Brookbank, J. W. (1984).** Karyogamy and meiosis in an *Amblyospora* sp (Microspora) in the mosquito *Culex salinarius*. *J Invertebr Pathol*, pp. 3-11.
- Hazard, E. I., Andreadis, T. G., Joslyn, D. J. & Ellis, E. A. (1979).** Meiosis and its implication in the life cycle of *Amblyospora* and *Parathelohania* (Microspora). *J Parasitol* **65**, 117-122.
- Heard E. 2004.** Recent advances in X-chromosome inactivation. *Curr Opin Cell Biol.* 16(3):247-55.
- Henikoff S, Ahmad K, Platero JS, van Steensel B. 2000.** Heterochromatic deposition of centromeric histone H3-like proteins. *Proc Natl Acad Sci U S A.* 97(2):716-21.
- Henikoff S, Matzke MA. 1997.** Exploring and explaining epigenetic effects. *Trends Genet.* ;13(8):293-5.
- Henson JD, Neumann AA, Yeager TR, Reddel RR. 2002.** Alternative lengthening of telomeres in mammalian cells. *Oncogene.* 21(4):598-610.
- Heus JJ, Zonneveld BJ, de Steensma HY, van den Berg JA. 1993.** The consensus sequence of *Kluyveromyces lactis* centromeres shows homology to functional centromeric DNA from *Saccharomyces cerevisiae*. *Mol Gen Genet.* 236(2-3):355-62.
- Heus JJ, Zonneveld BJ, Steensma HY, Van den Berg JA. 1994.** Mutational analysis of centromeric DNA elements of *Kluyveromyces lactis* and their role in determining the species specificity of the highly homologous centromeres from *K. lactis* and *Saccharomyces cerevisiae*. *Mol Gen Genet.* 243(3):325-33.
- Hinterberg K, Mattei D, Wellems TE, Scherf A. 1994.** Interchromosomal exchange of a large subtelomeric segment in a *Plasmodium falciparum* cross. *EMBO J.* 13(17):4174-80.
- Hirano T, Kobayashi R, Hirano M. 1997.** Condensins, chromosome condensation protein complexes containing XCAP-C, XCAP-E and a *Xenopus* homolog of the *Drosophila* Barren protein. *Cell.*89(4):511-21.
- Hirt, R. P., Healy, B., Vossbrinck, C. R., Canning, E. U. & Embley, T. M. (1997).** A mitochondrial Hsp70 orthologue in *Vairimorpha necatrix*: molecular evidence that microsporidia once contained mitochondria. *Curr Biol* **7**, 995-8.

- Hong ST, Steele PE, Cushion MT, Walzer PD, Stringer SL, Stringer JR. 1990.** *Pneumocystis carinii* karyotypes.
J Clin Microbiol. 28(8):1785-95.
- Horrocks P, Pinches R, Christodoulou Z, Kyes SA, Newbold CI. 2004.** Variable var transition rates underlie antigenic variation in malaria.
Proc Natl Acad Sci U S A.;101(30):11129-34. Epub 2004 Jul 15.
- Howard RJ, Barnwell JW, Kao V. 1983.** Antigenic variation of *Plasmodium knowlesi* malaria: identification of the variant antigen on infected erythrocytes.
Proc Natl Acad Sci U S A.;80(13):4129-33.
- Hoyer LL. 2001.** The ALS gene family of *Candida albicans*.
Trends Microbiol. ;9(4):176-80.
- Hughes AL, Friedman R. 2003.** Parallel evolution by gene duplication in the genomes of two unicellular fungi.
Genome Res. 13(6A):1259-64.
- Ishiara, R. & Hayashi, Y. (1968).** Some properties of ribosomes from sporoplasm of *Nosema bombycis*.
J Invertebr Pathol **11**, 377-385.
- Ivens AC, Peacock CS, Worthey EA, Murphy L et al. 2005.** The genome of the kinetoplastid parasite, *Leishmania major*.
Science. 309(5733):436-42.
- Iwano H., and Ishira R. 1991.** Dimorphism of spores of *Nosema spp.* In cultured cell.
J. Invertebr. Pathol. 57: 211-219.
- James TY, Kauff F, Schoch CL, et al. 2006.** Reconstructing the early evolution of Fungi using a six-gene phylogeny.
Nature. 443(7113):818-22.
- Jefferies D, Tebabi P, Pays E. 1991.** Transient activity assays of the *Trypanosoma brucei* variant surface glycoprotein gene promoter: control of gene expression at the posttranscriptional level.
Mol Cell Biol. 11(1):338-43.
- Jessberger R. 2002.** The many functions of SMC proteins in chromosome dynamics.
Nat Rev Mol Cell Biol. 3(10):767-78.
- Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT, Davis RW, Scherer S. 2004.** The diploid genome sequence of *Candida albicans*.
Proc Natl Acad Sci U S A. 101(19):7329-34. Epub 2004 May 3.
- Katinka MD, Duprat S, Cornillot E, Metenier G, Thomarat F, Prensier G, Barbe V, Peyretailade E, Brottier P, Wincker P, Delbac F, El Alaoui H, Peyret P, Saurin W, Gouy M, Weissenbach J, Vivares CP. 2001.** Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*.
Nature.;414(6862):450-3.
- Kawakami, Y., Inoue, T., Ito, K., Kitamizu, K., Hanawa, C., Ando, T., Iwano, H. & Ishihara, R. (1994).** Identification of a chromosome harboring the small subunit ribosomal RNA gene of *Nosema bombycis*.
J Invertebr Pathol **64**, 147-8.
- Keeling PJ, Doolittle WF. 1996.** Alpha-tubulin from early-diverging eukaryotic lineages and the evolution of the tubulin family.
Mol Biol Evol.;13(10):1297-305.
- Keeling PJ, Slamovits CH. 2004.** Simplicity and complexity of microsporidian genomes.
Eukaryot Cell. 3(6):1363-9.

- Kemp DJ, Thompson JK, Walliker D, Corcoran LM. 1987.** Molecular karyotype of *Plasmodium falciparum*: conserved linkage groups and expendable histidine-rich protein genes.
Proc Natl Acad Sci U S A. 84(21):7672-6.
- Kerry S. Bloom and John Carbon. 1982.** Yeast centromere DNA is in a unique and highly ordered structure in chromosomes and small circular minichromosomes.
Cell. 29(2):305-317
- Kimura K, Hirano T. 1997.** ATP-dependent positive supercoiling of DNA by 13S condensin: a biochemical implication for chromosome condensation.
Cell. 90(4):625-34.
- Kirk KE, Blackburn EH. 1995.** An unusual sequence arrangement in the telomeres of the germ-line micronucleus in *Tetrahymena thermophila*.
Genes Dev. 9(1):59-71.
- Kitada K, Yamaguchi E, Hamada K, Arisawa M. 1997.** Structural analysis of a *Candida glabrata* centromere and its functional homology to the *Saccharomyces cerevisiae* centromere.
Curr Genet. 31(2):122-7.
- Klein F, Laroche T, Cardenas ME, Hofmann JF, Schweizer D, Gasser SM. 1992.** Localization of RAP1 and topoisomerase II in nuclei and meiotic chromosomes of yeast.
J Cell Biol. 117(5):935-48.
- Kobryn K, Chaconas G. 2001.** The circle is broken: telomere resolution in linear replicons.
Curr Opin Microbiol. 4(5):558-64.
- Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, Makarova KS, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Rogozin IB, Smirnov S, Sorokin AV, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2004.** A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes.
Genome Biol.;5(2):R7. Epub 2004 Jan 15.
- Kooter JM, Borst P. 1984.** Alpha-amanitin-insensitive transcription of variant surface glycoprotein genes provides further evidence for discontinuous transcription in trypanosomes.
Nucleic Acids Res. 12(24):9457-72.
- Kornberg RD. 1974.** Chromatin structure: a repeating unit of histones and DNA.
Science 184(139):868-71.
- Kovacs JA, Halpern JL, Swan JC, Moss J, Parrillo JE, Masur H. 1988.** Identification of antigens and antibodies specific for *Pneumocystis carinii*.
J Immunol.;140(6):2023-31.
- Kyes S, Horrocks P, Newbold C. 2001.** Antigenic variation at the infected red cell surface in malaria.
Annu Rev Microbiol. 55:673-707.
- Kyes SA, Christodoulou Z, Raza A, Horrocks P, Pinches R, Rowe JA, Newbold CI. 2003.** A well-conserved *Plasmodium falciparum* var gene shows an unusual stage-specific transcript pattern.
Mol Microbiol. 48(5):1339-48.
- Kyrion G, Liu K, Liu C, Lustig AJ. 1993.** RAP1 and telomere structure regulate telomere position effects in *Saccharomyces cerevisiae*.
Genes Dev. 1993 Jul;7(7A):1146-59.
- Lachner M, O'Carroll D, Rea S, Mechtler K, Jenuwein T. 2001.** Methylation of histone H3 lysine 9 creates a binding site for HP1 proteins.
Nature. 410(6824):116-20.

- Ladurner AG. 2003.** Inactivating chromosomes: a macro domain that minimizes transcription.
Mol Cell. 12(1):1-3.
- Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K et al. International Human Genome Sequencing Consortium. 2001.** Initial sequencing and analysis of the human genome.
Nature 409(6822):860-921.
- Lanzer M, de Bruin D, Wertheimer SP, Ravetch JV. 1994.** Organization of chromosomes in *Plasmodium falciparum*: a model for generating karyotypic diversity.
Parasitol Today. 10(3):114-7.
- Larsson R. 1986.** Ultrastructure, function and classification of microsporidia. In: *Progress in Protistology*, edited by J. O. Corliss and D. J. Patterson (Biopress, Ltd, Bristol), p. 325-390.
- Lanzer M, Fischer K, Le Blancq SM. 1995.** Parasitism and chromosome dynamics in protozoan parasites: is there a connection?
Mol Biochem Parasitol. 70(1-2):1-8.
- Laroche T, Martin SG, Gotta M, Gorham HC, Pryde FE, Louis EJ, Gasser SM. 1998.** Mutation of yeast Ku genes disrupts the subnuclear organization of telomeres.
Curr Biol. 8(11):653-6.
- Larsson R. 1986.** Ultrastructure, function and classification of microsporidia. In: *Progress in Protistology*, edited by J. O. Corliss and D. J. Patterson (Biopress, Ltd, Bristol), p. 325-390.
- Lavstsen T, Salanti A, Jensen AT, Arnot DE, Theander TG. 2003.** Sub-grouping of *Plasmodium falciparum* 3D7 var genes based on sequence analysis of coding and non-coding regions.
Malar J.;2:27. Epub 2003 Sep 10.
- Lehker MW, Alderete JF. 1999.** Resolution of six chromosomes of *Trichomonas vaginalis* and conservation of size and number among isolates.
J Parasitol. 85(5):976-9.
- Lei M, Podell ER, Cech TR. 2004.** Structure of human POT1 bound to telomeric single-stranded DNA provides a model for chromosome end-protection.
Nat Struct Mol Biol. 11(12):1223-9. Epub 2004 Nov 21.
- Levaditi C, Nicolau S, Schoen R. 1923.** L'agent etiologique de l'encéphalite epizootique du lapin (*Encephalitozoon cuniculi*).
C.R. *Soc Biol Paris* 89:984-986.
- Li B, Oestreich S, de Lange T. 2000.** Identification of human Rap1: implications for telomere evolution.
Cell. 101(5):471-83.
- Li YC, Cheng TH, Gartenberg MR. 2001.** Establishment of transcriptional silencing in the absence of DNA replication.
Science. 2001 Jan 26;291(5504):650-3.
- Lingner J, Hughes TR, Shevchenko A, Mann M, Lundblad V, Cech TR. 1997.** Reverse transcriptase motifs in the catalytic subunit of telomerase.
Science. 276(5312):561-7.
- Linke MJ, Sunkin SM, Andrews RP, Stringer JR, Walzer PD. 1998.** Expression, structure, and location of epitopes of the major surface glycoprotein of *Pneumocystis carinii* f. sp. *carinii*.
Clin Diagn Lab Immunol.;5(1):50-7.
- Loayza D, De Lange T. 2003.** POT1 as a terminal transducer of TRF1 telomere length control.
Nature. 423(6943):1013-8.

- Loftus B, Anderson I, Davies R, 2005.** The genome of the protist parasite *Entamoeba histolytica*. *Nature*. 433(7028):865-8.
- Lom J, Nilsen F. 2003.** Fish microsporidia: fine structural diversity and phylogeny. *Int J Parasitol.* ;33(2):107-27.
- Losada A, Hirano M, Hirano T. 1998.** Identification of Xenopus SMC protein complexes required for sister chromatid cohesion. *Genes Dev.* 12(13):1986-97.
- Losada A, Hirano T. 2001.** Shaping the metaphase chromosome: coordination of cohesion and condensation. *Bioessays*. 23(10):924-35.
- Loubes C. 1979.** [Meiosis in Microsporidia: effects on biological cycles] *J Protozool.*;26(2):200-8.
- Louis EJ, Haber JE. 1990.** Mitotic recombination among subtelomeric Y' repeats in *Saccharomyces cerevisiae*. *Genetics*. 124(3):547-59.
- Louis EJ, Naumova ES, Lee A, Naumov G, Haber JE. 1994.** The chromosome end in yeast: its mosaic nature and influence on recombinational dynamics. *Genetics*. 136(3):789-802.
- Louis EJ. 1995.** The chromosome ends of *Saccharomyces cerevisiae*. *Yeast*. 11(16):1553-73.
- Lowe J, Cordell SC, van den Ent F. 2001.** Crystal structure of the SMC head domain: an ABC ATPase with 900 residues antiparallel coiled-coil inserted. *J Mol Biol.* 306(1):25-35.
- Lowman PM, Takvorian PM, Cali A. 2000.** The effects of elevated temperatures and various time-temperature combinations on the development of *Brachiola (Nosema) algerae* N. Comb. in mammalian cell culture. *J Eukaryot Microbiol.* ;47(3):221-34.
- Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. 1997.** Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389(6648):251-60.
- Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. 2004.** The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res.*;32(Database issue):D235-9.
- Malone, L. A. & McIvor, C. A. (1993).** Pulse-Field Gel Electrophoresis of DNA from four Microsporidian Isolates. *Journal of Invertebrate Pathology* **61**, 203-205.
- Mansour L, Cheikali C, Desautais P, Coulon JP, Daubin J, Hassine OK, Vivares CP, Jeanjean J, Cornillot E. (2004).** Description of an ultrathin multiwire proportional chamber-based detector and application to the characterization of the *Spraguea lophii* (Microspora) two-dimensional genome fingerprint. *Electrophoresis* 25(20):3365-77.
- Marcand S, Wotton D, Gilson E, Shore D. 1997.** Rap1p and telomere length regulation in yeast. *Ciba Found Symp.* 211:76-93; discussion 93-103.
- Marti M, Good RT, Rug M, Knuepfer E, Cowman AF. 2004.** Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science*. ;306(5703):1930-3.

- Martin CL, Wong A, Gross A, Chung J, Fantes JA, Ledbetter DH. 2002.** The evolutionary origin of human subtelomeric homologies- or where the ends begin.
Am J Hum Genet. 70(4):972-84. Epub 2002 Mar 1.
- Martin SG, Laroche T, Suka N, Grunstein M, Gasser SM. 1999.** Relocalization of telomeric Ku and SIR proteins in response to DNA strand breaks in yeast.
Cell. 97(5):621-33.
- Mason JM, Biessmann H. 1995.** The unusual telomeres of *Drosophila*.
Trends Genet. 11(2):58-62.
- Mathis A, Tanner I, Weber R, Deplazes P. 1999.** Genetic and phenotypic intraspecific variation in the microsporidian *Encephalitozoon hellem*.
Int J Parasitol. ;29(5):767-70.
- Mathis A, Weber R, Deplazes P. 2005.** Zoonotic potential of the microsporidia.
Clin Microbiol Rev.;18(3):423-45.
- Maxam AM, Gilbert W. 1992.** A new method for sequencing DNA. 1977.
Biotechnology. 24:99-103.
- McClelland M, Jones R, Patel Y, Nelson M. 1987.** Restriction endonucleases for pulsed field mapping of bacterial genomes.
Nucleic Acids Res. 1987 Aug 11;15(15):5985-6005.
- McClintock B. 1939.** The behavior in successive nuclear divisions of a chromosome broken at meiosis.
Proc Natl Acad Sci USA 25: 405-16.
- McCulloch R, Rudenko G, Borst P. 1997.** Gene conversions mediating antigenic variation in *Trypanosoma brucei* can occur in variant surface glycoprotein expression sites lacking 70-base-pair repeat sequences.
Mol Cell Biol. ;17(2):833-43.
- Mefford HC, Trask BJ. 2002.** The complex structure and dynamic evolution of human subtelomeres.
Nat Rev Genet. 3(2):91-102.
- Melby TE, Ciampaglio CN, Briscoe G, Erickson HP. 1998.** The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: long, antiparallel coiled coils, folded at a flexible hinge.
J Cell Biol. 142(6):1595-604.
- Melville SE, Leech V, Navarro M, Cross GA. 2000.** The molecular karyotype of the megabase chromosomes of *Trypanosoma brucei* stock 427.
Mol Biochem Parasitol. 111(2):261-73.
- Meneghini MD, Wu M, Madhani HD. 2003.** Conserved histone variant H2A.Z protects euchromatin from the ectopic spread of silent heterochromatin.
Cell 112(5):725-36.
- Michaelis C, Ciosk R, Nasmyth K. 1997.** Cohesins: chromosomal proteins that prevent premature separation of sister chromatids.
Cell. 91(1):35-45.
- Miklos I, Varga T, Nagy A, Sipiczki M. 1997.** Genome instability and chromosomal rearrangements in a heterothallic wine yeast.
J Basic Microbiol. 37(5):345-54.
- Miller LH, Good MF, Milon G. 1994.** Malaria pathogenesis.
Science. 264(5167):1878-83.

- Mittleider D, Green LC, Mann VH, Michael SF, Didier ES, Brindley PJ. 2002.** Sequence survey of the genome of the opportunistic microsporidian pathogen, *Vittaforma corneae*.
J Eukaryot Microbiol.;49(5):393-401.
- Moazed D. 2001.** Common themes in mechanisms of gene silencing.
Mol Cell. 8(3):489-98.
- Moazed D. 2001.** Enzymatic activities of Sir2 and chromatin silencing.
Curr Opin Cell Biol. ;13(2):232-8.
- Muller HJ. 1938.** The remaking of chromosomes.
The Collecting Net 13: 181-95, 198.
- Munderloh, U. G., Kurtti, T. J. & Ross, S. E. (1990).** Electrophoretic characterization of chromosomal DNA from two microsporidia.
J Invertebr Pathol **56**, 243-8.
- Murray AW, Szostak JW.1983.** Construction of artificial chromosomes in yeast.
Nature. 305(5931):189-93.
- Murti KG, Prescott DM. 1999.** Telomeres of polytene chromosomes in a ciliated protozoan terminate in duplex DNA loops.
Proc Natl Acad Sci U S A. 1999 Dec 7;96(25):14436-9.
- Nakamura Y, Leppert M, O'Connell P, Wolff R, Holm T, Culver M, Martin C, Fujimoto E, Hoff M, Kumlin E, et al. 1987.** Variable number of tandem repeat (VNTR) markers for human gene mapping.
Science. 235(4796):1616-22.
- Nash T. 1992.** Surface antigen variability and variation in *Giardia lamblia*.
Parasitol Today. 8(7):229-34.
- Nash TE, Mowatt MR. 1992.** Characterization of a *Giardia lamblia* variant-specific surface protein (VSP) gene from isolate GS/M and estimation of the VSP gene repertoire size.
Mol Biochem Parasitol.;51(2):219-27.
- Nassonova E, Cornillot E, Metenier G, Agafonova N, Kudryavtsev B, Skarlato S, Vivares CP. 2005.** Chromosomal composition of the genome in the monomorphic diplokaryotic microsporidium *Paranosema grylli*: analysis by two-dimensional pulsed-field gel electrophoresis.
Folia Parasitol (Praha). ;52(1-2):145-57.
- Natale DA, Schubert AE, Kowalski D. 1992.** DNA helical stability accounts for mutational defects in a yeast replication origin.
Proc Natl Acad Sci U S A. 89(7):2654-8.
- Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin AG, Murzina NV, Laue ED. 2002.** Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9.
Nature. 416(6876):103-7.
- Novacek, M.J. and M.A. Norell. 1989.** Nuclear DNA content in bats and other organisms: implications and unanswered questions.
Trends in Ecology and Evolution 10(4):285-286.
- Orenstein J M. 1991.** Microsporidiosis in the acquired immunodeficiency Syndrome.
J Parasitol. 77:843-864

- Paques F, Haber JE. 1999.** Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*.
Microbiol Mol Biol Rev. 63(2):349-404.
- Patel PK, Arcangioli B, Baker SP, Bensimon A, Rhind N. 2006.** DNA replication origins fire stochastically in fission yeast.
Mol Biol Cell. 17(1):308-16.
- Pays E, Coquelet H, Tebabi P, Pays A, Jefferies D, Steinert M, Koenig E, Williams RO, Roditi I. 1990.** *Trypanosoma brucei*: constitutive activity of the VSG and procyclin gene promoters.
EMBO J. 9(10):3145-51.
- Pays E, Lheureux M, Steinert M. 1981.** The expression-linked copy of a surface antigen gene in *Trypanosoma* is probably the one transcribed.
Nature. ;292(5820):265-7.
- Pays E, Lips S, Nolan D, Vanhamme L, Perez-Morga D. 2001.** The VSG expression sites of *Trypanosoma brucei*: multipurpose tools for the adaptation of the parasite to mammalian hosts.
Mol Biochem Parasitol. ;114(1):1-16.
- Pays E, Nolan DP. 1998.** Expression and function of surface proteins in *Trypanosoma brucei*.
Mol Biochem Parasitol. 91(1):3-36.
- Peyret P, Katinka MD, Duprat S, Duffieux F, Barbe V, Barbazanges M, Weissenbach J, Saurin W, Vivares CP. 2005.** Sequence and analysis of chromosome I of the amitochondriate intracellular parasite *Encephalitozoon cuniculi* (Microspora).
Genome Res.;11(2):198-207.
- Peyretailade, E., Biderre, C., Peyret, P., Duffieux, F., Méténier, G., Gouy, M., Michot, B. & Vivarès, C. P. (1998b).** Microsporidian *Encephalitozoon cuniculi*, a unicellular eukaryote with an unusual chromosomal dispersion of ribosomal genes and a LSU rRNA reduced to the universal core.
Nucleic Acids Res **26**, 3513-20.
- Peyretailade, E., Broussolle, V., Peyret, P., Méténier, G., Gouy, M. & Vivarès, C. P. (1998a).** Microsporidia, amitochondrial protists, possess a 70-kDa heat shock protein gene of mitochondrial evolutionary origin.
Mol Biol Evol **15**, 683-9.
- Pietrasanta LI, Thrower D, Hsieh W, Rao S, Stemmann O, Lechner J, Carbon J, Hansma H. 1999.** Probing the *Saccharomyces cerevisiae* centromeric DNA (CEN DNA)-binding factor 3(CBF3) kinetochore complex by using atomic force microscopy.
Proc Natl Acad Sci U S A. 96(7):3757-62.
- Pluta AF, Mackay AM, Ainsztein AM, Goldberg IG, Earnshaw WC. 1995.** The centromere: hub of chromosomal activities.
Science. 270(5242):1591-4.
- Pottratz ST, Martin WJ 2nd. 1990.** Mechanism of *Pneumocystis carinii* attachment to cultured rat alveolar macrophages.
J Clin Invest.;86(5):1678-83.
- Prensier G, Slomianny C. 1986.** The karyotype of *Plasmodium falciparum* determined by ultrastructural serial sectioning and 3D reconstruction.
J Parasitol. 72(5):731-6.
- Raikov, I. B. (1994).** The diversity of forms of mitosis in Protozoa : a comparative review. **30**, 253-269.

- Ralph SA, Bischoff E, Mattei D, Sismeiro O, Dillies MA, Guigon G, Coppee JY, David PH, Scherf A. 2005.** Transcriptome analysis of antigenic variation in *Plasmodium falciparum*—var silencing is not dependent on antisense RNA. *Genome Biol.* 2005;6(11):R93.
- Rao H, Stillman B. 1995.** The origin recognition complex interacts with a bipartite DNA binding site within yeast replicators. *Proc Natl Acad Sci U S A.* 92(6):2224-8.
- Ridgway P, Almouzni G. 2001.** Chromatin assembly and organization. *J Cell Sci.* 114(Pt 15):2711-2.
- Riles L, Dutchik JE, Baktha A, McCauley BK, Thayer EC, Leckie MP, Braden VV, Depke JE, Olson MV. 1993.** Physical maps of the six smallest chromosomes of *Saccharomyces cerevisiae* at a resolution of 2.6 kilobase pairs. *Genetics.* 134(1):81-150.
- Riou JF, Gomez D, Mergny JL, Guittat L, Paterski R, Chenais B, Morjani H, Trentesaux C. 2005.** [Regulation of telomeres length: making the telomeres accessible?] *Bull Cancer.* 92(1):13-22.
- Rogakou EP, Pilch DR, Orr AH, Ivanova VS, Bonner WM. 1998.** DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem.* 273(10):5858-68.
- Rudenko G, Blundell PA, Dirks-Mulder A, Kieft R, Borst P. 1995.** A ribosomal DNA promoter replacing the promoter of a telomeric VSG gene expression site can be efficiently switched on and off in *T. brucei*. *Cell.* 83(4):547-53.
- Rudenko G, Cross M, Borst P. 1998.** Changing the end: antigenic variation orchestrated at the telomeres of African trypanosomes. *Trends Microbiol.* 6(3):113-6.
- Rustchenko EP, Curran TM, Sherman F. 1993.** Variations in the number of ribosomal DNA units in morphological mutants and normal strains of *Candida albicans* and in normal strains of *Saccharomyces cerevisiae*. *J Bacteriol.* 175(22):7189-99.
- Salanti A, Staalsoe T, Lavstsen T, Jensen AT, Sowa MP, Arnot DE, Hviid L, Theander TG. 2003.** Selective upregulation of a single distinctly structured var gene in chondroitin sulphate A-adhering *Plasmodium falciparum* involved in pregnancy-associated malaria. *Mol Microbiol.* 49(1):179-91.
- Sandell LL, Gottschling DE, Zakian VA. 1994.** Transcription of a yeast telomere alleviates telomere position effect without affecting chromosome stability. *Proc Natl Acad Sci U S A.* 91(25):12061-5.
- Sanger F, Nicklen S, Coulson AR. 1992.** DNA sequencing with chain-terminating inhibitors. 1977. *Biotechnology.* 24:104-8.
- Scherf A, Carter R, Petersen C, Alano P, Nelson R, Aikawa M, Mattei D, Pereira da Silva L, Leech J. 1992.** Gene inactivation of Pf11-1 of *Plasmodium falciparum* by chromosome breakage and healing: identification of a gametocyte-specific protein with a potential role in gametogenesis. *EMBO J.* 11(6):2293-301.
- Scherf A, Figueiredo LM, Freitas-Junior LH. 2001.** *Plasmodium* telomeres: a pathogen's perspective. *Curr Opin Microbiol.* 4(4):409-14.

- Scherf A, Hernandez-Rivas R, Buffet P, Bottius E, Benatar C, Pouvelle B, Gysin J, Lanzer M. 1998.** Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in *Plasmodium falciparum*. *EMBO J.*;17(18):5418-26.
- Schwob E. 2004.** Flexibility and governance in eukaryotic DNA replication. *Curr Opin Microbiol.* 7(6):680-90.
- Shore D. 2001.** Telomeric chromatin: replicating and wrapping up chromosome ends. *Curr Opin Genet Dev.* 11(2):189-98.
- Sibley LD, Boothroyd JC. 1992.** Construction of a molecular karyotype for *Toxoplasma gondii*. *Mol Biochem Parasitol.* 51(2):291-300.
- Simpson RT. 1978.** Structure of the chromatosome, a chromatin particle containing 160 base pairs of DNA and all the histones. *Biochemistry* 17(25):5524-31.
- Slamovits CH, Fast NM, Law JS, Keeling PJ. 2004.** Genome compaction and stability in microsporidian intracellular parasites. *Curr Biol.* ;14(10):891-6.
- Smith S, Giriat I, Schmitt A, de Lange T. 1998.** Tankyrase, a poly(ADP-ribose) polymerase at human telomeres. *Science.* 282(5393):1484-7.
- Smith S, de Lange T. 2000.** Tankyrase promotes telomere elongation in human cells. *Curr Biol.* 10(20):1299-302.
- Smogorzewska A, van Steensel B, Bianchi A, Oelmann S, Schaefer MR, Schnapp G, de Lange T. 2000.** Control of human telomere length by TRF1 and TRF2. *Mol Cell Biol.* 20(5):1659-68.
- Sobottka I, Albrecht H, Visvesvara GS, Pieniazek NJ, Deplazes P, Schwartz DA, Laufs R, Elsner HA. 1999.** Inter- and intra-species karyotype variations among microsporidia of the genus *Encephalitozoon* as determined by pulsed-field gel electrophoresis. *Scand J Infect Dis.*31(6):555-8.
- Sprague, V. (1977).** Classification and phylogeny of the microsporidian. In *Comparative Pathobiology*, pp. 1-30. Edited by L. A. B. a. T. C. Cheng. New-York and London: Plenum Press.
- Staalsoe T, Hamad AA, Hviid L, Elhassan IM, Arnot DE, Theander TG. 2002.** In vivo switching between variant surface antigens in human *Plasmodium falciparum* infection. *J Infect Dis.*;186(5):719-22. Epub 2002 Aug 5.
- Stein LD. 2004.** Human genome: end of the beginning. *Nature* 21;431(7011):915-6.
- Steiner NC, Hahnenberger KM, Clarke L. 1993.** Centromeres of the fission yeast *Schizosaccharomyces pombe* are highly variable genetic loci. *Mol Cell Biol.* 13(8):4578-87.
- Sternberg N. 1990.** Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. *Proc Natl Acad Sci U S A.* 87(1):103-7.

- Street DA. 1994.** Analysis of *Nosema locustae* (Microsporidia: nosematidae) chromosomal DNA with pulsed-field gel electrophoresis;
J. Invertebr. Pathol. 63: 301-303.
- Stringer JR, Keely SP. 2001.** Genetics of surface antigen expression in *Pneumocystis carinii*.
Infect Immun. ;69(2):627-39.
- Stringer JR. 2002.** *Pneumocystis*.
Int J Med Microbiol. 2002 Oct;292(5-6):391-404.
- Su X, Wellems TE. 1996.** Toward a high-resolution *Plasmodium falciparum* linkage map: polymorphic markers from hundreds of simple sequence repeats.
Genomics. 33(3):430-44.
- Su XZ, Heatwole VM, Wertheimer SP, Guinet F, Herrfeldt JA, Peterson DS, Ravetch JA, Wellems TE. 1995.** The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes.
Cell.;82(1):89-100.
- Sunkin SM, Stringer JR. 1996.** Translocation of surface antigen genes to a unique telomeric expression site in *Pneumocystis carinii*.
Mol Microbiol.;19(2):283-95.
- Sunkin SM, Stringer SL, Stringer JR. 1994.** A tandem repeat of rat-derived *Pneumocystis carinii* genes encoding the major surface glycoprotein.
J Eukaryot Microbiol.;41(3):292-300.
- Sutani T, Yuasa T, Tomonaga T, Dohmae N, Takio K, Yanagida M. 1999.** Fission yeast condensin complex: essential roles of non-SMC subunits for condensation and Cdc2 phosphorylation of Cut3/SMC4.
Genes Dev. 13(17):2271-83.
- Suto RK, Clarkson MJ, Tremethick DJ, Luger K. 2000.** Crystal structure of a nucleosome core particle containing the variant histone H2A.Z.
Nat Struct Biol. 7(12):1121-4.
- Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW. 1983.** The double-strand-break repair model for recombination.
Cell. 33(1):25-35.
- Taupin V, Garenaux E, Mazet M, Maes E, Denise H, Prensier G, Vivares CP, Guerardel Y, Metenier G. 2006.** Major O-glycans in the spores of two microsporidian parasites are represented by unbranched manno-oligosaccharides containing {alpha}-1,2 linkages.
Glycobiology. [Epub ahead of print]
- Theobald DL, Wuttke DS. 2004.** Prediction of multiple tandem OB-fold domains in telomere end-binding proteins Pot1 and Cdc13.
Structure. 12(10):1877-9.
- Thomarat F, Vivares CP, Gouy M. 2004.** Phylogenetic analysis of the complete genome sequence of *Encephalitozoon cuniculi* supports the fungal origin of microsporidia and reveals a high frequency of fast-evolving genes.
J Mol Evol. ;59(6):780-91.
- Trammer T, Chioralia G, Maier WA, Seitz HM. 1999.** In vitro replication of *Nosema algerae* (Microsporidia), a parasite of anopheline mosquitoes, in human cells above 36 degrees C.
J Eukaryot Microbiol. 46(5):464-8.
- Upcroft JA, Delgadillo-Correa MG, Dunne RL, Sturm AW, Johnson PJ, Upcroft P. 2006.** Genotyping *Trichomonas vaginalis*.
Int J Parasitol. 36(7):821-8.

- Van de Peer Y, Baldauf SL, Doolittle WF, Meyer A. 2000.** An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances.
J Mol Evol. ;51(6):565-76.
- van de Putte P, Goosen N. 1992.** DNA inversions in phages and bacteria.
Trends Genet. ;8(12):457-62.
- Van der Ploeg LH, Smith CL, Polvere RI, Gottesdiener KM. 1989.** Improved separation of chromosome-sized DNA from *Trypanosoma brucei*, stock 427-60.
Nucleic Acids Res. 17(8):3217-27.
- van Gent DC, Hoeijmakers JH, Kanaar R. 2001.** Chromosomal stability and the DNA double-stranded break connection.
Nat Rev Genet. 2(3):196-206.
- van Steensel B, de Lange T. 1997.** Control of telomere length by the human telomeric protein TRF1.
Nature. 385(6618):740-3.
- van Steensel B, Smogorzewska A, de Lange T. 1998.** TRF2 protects human telomeres from end-to-end fusions.
Cell. 92(3):401-13.
- Vanhamme L, Berberof M, Le Ray D, Pays E. 1995.** Stimuli of differentiation regulate RNA elongation in the transcription units for the major stage-specific antigens of *Trypanosoma brucei*.
Nucleic Acids Res. ;23(11):1862-9.
- Vanhamme L, Pays E. 1995.** Control of gene expression in trypanosomes.
Microbiol Rev. ;59(2):223-40.
- Vavra J., Larsson JJ. 1999.** Structure of the Microsporidia. In: The Microsporidia and Microsporidiosis, edited by M. Wittner and L. Weiss (American Society for microbiology, Washington, DC), p. 7-84.
- Vavra J., Horak A., Modry D., Lukes J. and Koudela B. 2006.** *Trachipleistophora extenrec* n. sp. a New Microsporidian (Fungi: Microsporidia) Infecting Mammals.
J. Eukaryot. Microbiol. 0(0), 10.1111/j.1550-7408.
- Vega-Palas MA, Martin-Figueroa E, Florencio FJ. 2000.** Telomeric silencing of a natural subtelomeric gene.
Mol Gen Genet. 263(2):287-91.
- Verstrepen KJ, Klis FM. 2006.** Flocculation, adhesion and biofilm formation in yeasts.
Mol Microbiol. ;60(1):5-15.
- Visvesvara GS, Belloso M, Moura H, Da Silva AJ, Moura IN, Leitch GJ, Schwartz DA, Chevez-Barrios P, Wallace S, Pieniazek NJ, Goosey JD. 1999.** Isolation of *Nosema algerae* from the cornea of an immunocompetent patient.
J Eukaryot Microbiol. 46(5):10S.
- Vivarès C.P., Sprague V. 1979.** The fine structure of *Ameson pulvis* (Microspora, Microsporidia) and its implication regarding classification and chromosome cycle.
J. Invertbr. Pathol. 33, 40-52.
- Vivarès CP, Gouy M, Thomarat F, Metenier G. 2002.** Functional and evolutionary analysis of a eukaryotic parasitic genome.
Curr Opin Microbiol. 2002 Oct;5(5):499-505.
- Vivarès, C. P., Bouix, G. & Manier, J. F. (1977).** *Ormiersia carcini* gen. n., sp. n., microsporidian of the Mediterranean crab, *Carcinus mediterraneus* Czerniavsky, 1844: developmental cycle and ultrastructural study. *J Protozool* 24, 83-94.

- Vivarès CP, Méténier G. 2001.** The microsporidian *Encephalitozoon*.
Bioessays 23, 194-202.
- Voss TS, Healer J, Marty AJ, Duffy MF, Thompson JK, Beeson JG, Reeder JC, Crabb BS, Cowman AF. 2006.** A var gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria.
Nature. 2006 Feb 23;439(7079):1004-8.
- Voss TS, Kaestli M, Vogel D, Bopp S, Beck HP. 2003.** Identification of nuclear proteins that interact differentially with *Plasmodium falciparum* var gene promoters.
Mol Microbiol. 48(6):1593-607.
- Voss TS, Thompson JK, Waterkeyn J, Felger I, Weiss N, Cowman AF, Beck HP. 2000.** Genomic distribution and functional characterisation of two distinct and conserved *Plasmodium falciparum* var gene 5' flanking sequences.
Mol Biochem Parasitol. ;107(1):103-15.
- Vossbrinck CR, Andreadis TG, Vavra J, Becnel JJ. 2004.** Molecular phylogeny and evolution of mosquito parasitic Microsporidia (Microsporidia: Amblyosporidae).
J Eukaryot Microbiol.;51(1):88-95.
- Vossbrinck, C. R. & Woese, C. R. (1986).** Eukaryotic ribosomes that lack a 5.8S RNA.
Nature 320, 287-288.
- Vossbrinck, C. R., Maddox, J. V., Friedman, S., Debrunner-Vossbrinck, B. A. & Woese, C. R. (1987).** Ribosomal RNA sequence suggests microsporidia are extremely ancient eukaryotes.
Nature 326, 411-4.
- Wada M, Nakamura Y. 1996.** Unique telomeric expression site of major-surface-glycoprotein genes of *Pneumocystis carinii*.
DNA Res.;3(2):55-64.
- Weber JL. 1987.** Analysis of sequences from the extremely A + T-rich genome of *Plasmodium falciparum*.
Gene. 52(1):103-9.
- Weidner, E., Byrd, W., Scarborough, A., Pleshinger, J. & Sibley, D. (1984).** Microsporidian spore discharge and the transfer of polaroplast organelle membrane into plasma membrane.
J. Protozool 31, 195-198.
- Wiens GR, Sorger PK. 1998.** Centromeric chromatin and epigenetic effects in kinetochore assembly.
Cell. 93(3):313-6.
- Willhoeft U, Tannich E. 1999.** The electrophoretic karyotype of *Entamoeba histolytica*.
Mol Biochem Parasitol. 99(1):41-53.
- Williams BA, Hirt RP, Lucocq JM, Embley TM. 2000.** A mitochondrial remnant in the microsporidian *Trachipleistophora hominis*.
Nature. ;418(6900):865-9.
- Wolfe KH, Shields DC. 1997.** Molecular evidence for an ancient duplication of the entire yeast genome.
Nature. 387(6634):708-13.
- Wood V, Gwilliam R, Rajandream MA, et al. 2002.** The genome sequence of *Schizosaccharomyces pombe*.
Nature. 415(6874):871-80.
- Workmann JL, Kingston RE. 1998.** Alteration of nucleosome structure as a mechanism of transcriptional regulation.
Annu Rev Biochem. ;67:545-79.

- Wotton D, Shore D. 1997.** A novel Rap1p-interacting factor, Rif2p, cooperates with Rif1p to regulate telomere length in *Saccharomyces cerevisiae*.
Genes Dev. 11(6):748-60.
- Wright JH and Craighead EM. 1922.** Infectious motor paralysis in young rabbits.
J. Exp. Med. 36: 135-140
- Wyder MA, Rasch EM, Kaneshiro ES. 1998.** Quantitation of absolute *Pneumocystis carinii* nuclear DNA content. Trophic and cystic forms isolated from infected rat lungs are haploid organisms.
J Eukaryot Microbiol. 45(3):233-9.
- Wyrick JJ, Holstege FC, Jennings EG, Causton HC, Shore D, Grunstein M, Lander ES, Young RA. 1999.** Chromosomal landscape of nucleosome-dependent gene expression and silencing in yeast.
Nature. 402(6760):418-21.
- Xong HV, Vanhamme L, Chamekh M, Chimfwembe CE, Van Den Abbeele J, Pays A, Van Meirvenne N, Hamers R, De Baetselier P, Pays E. 1998.** A VSG expression site-associated gene confers resistance to human serum in *Trypanosoma rhodesiense*.
Cell. 95(6):839-46.
- Yang Y, Adam RD. 1994.** Allele-specific expression of a variant-specific surface protein (VSP) of *Giardia lamblia*.
Nucleic Acids Res. ;22(11):2102-8.
- Ye JZ, de Lange T. 2004.** TIN2 is a tankyrase 1 PARP modulator in the TRF1 telomere length control complex.
Nat Genet. 36(6):618-23.
- Ziegelbauer K, Overath P. 1993.** Organization of two invariant surface glycoproteins in the surface coat of *Trypanosoma brucei*.
Infect Immun. 61(11):4540-5.
- Zomerdijk JC, Kieft R, Shiels PG, Borst P. 1991.** Alpha-amanitin-resistant transcription units in trypanosomes: a comparison of promoter sequences for a VSG gene expression site and for the ribosomal RNA genes.
Nucleic Acids Res.;19(19):5153-8.